FOURTH EDITION

CONTEMPORARY PRACTICE IN CLINICAL CHEMISTRY

EDITED BY
WILLIAM CLARKE AND MARK A. MARZINKE

AACC | PRESS

Contemporary Practice in Clinical Chemistry

Contemporary Practice in Clinical Chemistry

Fourth Edition

Edited by

William Clarke

Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Mark A. Marzinke

Departments of Pathology and Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Published in cooperation with AACC

AACC | PRESS

ELSEVIER

ACADEMIC PRESS

An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

About AACC

Dedicated to achieving better health through laboratory medicine, AACC brings together more than 50,000 clinical laboratory professionals, physicians, research scientists, and business leaders from around the world focused on clinical chemistry, molecular diagnostics, mass spectrometry, translational medicine, lab management, and other areas of progressing laboratory science. Since 1948, AACC has worked to advance the common interests of the field, providing programs that advance scientific collaboration, knowledge, expertise, and innovation. For more information, visit www.aacc.org.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-815499-1

For Information on all Academic Press publications
visit our website at https://www.elsevier.com/books-and-journals

Publisher: Stacy Masucci
Acquisitions Editor: Tari K. Broderick
Editorial Project Manager: Pat Gonzalez
Production Project Manager: Omer Mukthar
Cover Designer: Mark Rogers

Typeset by MPS Limited, Chennai, India

Contents

List of contributors

Khosrow Adeli, Clinical Biochemistry, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada

David N. Alter, Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA, United States

Esther Babady, Clinical Microbiology Service, Department of Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, United States

Lorin M. Bachmann, Department of Pathology, Virginia Commonwealth University, Richmond, VA, United States

Shi Bai, Department of Pathology, UMass Memorial Medical School, Worcester, MA, United States

Darci R. Block, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, United States

Sean Campbell, Department of Pathology, Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, NY, United States

Allison B. Chambliss, Department of Pathology, Keck School of Medicine of USC, Los Angeles, CA, United States

Daniel W. Chan, Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD, United States

Alex C. Chin, Alberta Precision Laboratories and Department of Pathology and Laboratory Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

Robert H. Christenson, Department of Pathology, University of Maryland School of Medicine, Baltimore, MD, United States; Department of Medical and Research Technology, University of Maryland School of Medicine, Baltimore, MD, United States

William Clarke, Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD, United States

Jennifer M. Colby, Department of Pathology, Microbiology, and Immunology, Vanderbilt University School of Medicine, Nashville, TN, United States

Margery A. Connelly, Laboratory Corporation of America Holdings (LabCorp), NC, United States

Steven W. Cotten, Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Jayme L. Dahlin, Department of Pathology, Brigham and Women's Hospital, Boston, MA, United States; Harvard Medical School, Boston, MA, United States

Sophie J. Deharvengt, Department of Pathology and Laboratory Medicine, Dartmouth-Hitchcock Medical Center and Norris Cotton Cancer Center, Lebanon, NH, United States; Geisel School of Medicine at Dartmouth, Hanover, NH, United States

Edgard Delvin, Division of Gastroenterology, Hepatology and Nutrition, CHU Sainte Justine Research Centre, Montreal, QC, Canada; Department of Biochemistry, Faculty of Medicine, University of Montreal, Montreal, QC, Canada

Jane Dickerson, Department of Laboratory Medicine, University of Washington Medical Center, Seattle, WA, United States

Leslie J. Donato, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, United States

Paul D'Orazio, Instrumentation Laboratory, Bedford, MA, United States

D. Robert Dufour, Department of Pathology (Emeritus), George Washington University, Washington, DC, United States; Department of Pathology (Adjunct), Northwestern University Feinberg School of Medicine, Chicago, IL, United States

Show-Hong Duh, Department of Pathology, University of Maryland School of Medicine, Baltimore, MD, United States

Ashley Eckel, Department of Laboratory Medicine, University of Washington, Seattle, WA, United States

Shu-Ling Fan, Department of Pathology, UMass Memorial Medical School, Worcester, MA, United States

Elizabeth L. Frank, Department of Pathology, University of Utah School of Medicine, Salt Lake City, UT, United States; ARUP Laboratories, Inc., Salt Lake City, UT, United States

Deanna D.H. Franke, Technical Specialties, Atrium Health, Corporate Lab Services, NC, United States

Jonathan R. Genzen, Department of Pathology, University of Utah School of Medicine & ARUP Laboratories, Salt Lake City, UT, United States

Dina N. Greene, Washington Kaiser Permanente, Renton, WA, United States

David S. Hage, Department of Chemistry, University of Nebraska, Lincoln, NE, United States

Neil S. Harris, Department of Pathology, University of Florida, Gainesville, FL, United States

Matthew P.A. Henderson, Newborn Screening Ontario, The Children's Hospital of Eastern Ontario Department of Pediatrics, University of Ottawa, Ottawa, ON, Canada

Victoria Higgins, Clinical Biochemistry, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada

Daniel T. Holmes, Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada

Lesley A. Inker, Department of Medicine, Division of Nephrology, Tufts Medical Center, Boston, MA, United States

Brian R. Jackson, Department of Pathology, University of Utah, Salt Lake City, UT, United States; ARUP Laboratories, Salt Lake City, UT, United States

Paul J. Jannetto, Clinical and Forensic Toxicology Laboratory, Mayo Clinic, Rochester, MN, United States

Hou-Sung Jung, Department of Pathology and Laboratory Medicine, Dartmouth-Hitchcock Medical Center and Norris Cotton Cancer Center, Lebanon, NH, United States; Geisel School of Medicine at Dartmouth, Hanover, NH, United States

Stephen E. Kahn, Department of Pathology, Loyola University Medical Center, Maywood, IL, United States

Peter A. Kavsak, Department of Pathology and Molecular Medicine, McMaster University, Hamilton, ON, Canada; Juravinski Hospital and Cancer Centre, Hamilton Health Sciences, Hamilton, ON, Canada

Nichole Korpi-Steiner, Department of Pathology and Laboratory Medicine, University of North Carolina, Chapel Hill, NC, United States

Loralie J. Langman, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, United States

Stanley T. Leung, Incyte Diagnostics, Seattle, WA, United States

Emile Levy, Division of Gastroenterology, Hepatology and Nutrition, CHU Sainte-Justine Research Centre, Montreal, QC, Canada; Department of Nutrition, Faculty of Medicine, University of Montreal, Montreal, QC, Canada

Rachael Liesman, Clinical Microbiology, Department of Pathology & Laboratory Medicine, The University of Kansas Health System, Kansas City, KS, United States

David C. Lin, Department of Pathology, Ann & Robert H. Lurie Children's Hospital of Chicago, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

Christina M. Lockwood, Department of Laboratory Medicine, University of Washington, Seattle, WA, United States

Mark A. Marzinke, Departments of Pathology and Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Stephen R. Master, Division of Laboratory Medicine, Children's Hospital of Philadelphia, Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, United States

Christopher R. McCudden, Eastern Ontario Regional Laboratory Association, The Ottawa Hospital Department of Pathology and Laboratory Medicine, University of Ottawa, Ottawa, ON, Canada

Gwendolyn A. McMillin, Department of Pathology at the University of Utah School of Medicine, ARUP Laboratories, Salt Lake City, UT, United States

Jeffrey W. Meeusen, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, United States

Stacy E.F. Melanson, Department of Pathology, Brigham and Women's Hospital, Boston, MA, United States; Harvard Medical School, Boston, MA, United States

Anna E. Merrill, Department of Pathology, University of Iowa Hospitals and Clinics, Iowa City, IA, United States

W. Greg Miller, Department of Pathology, Virginia Commonwealth University, Richmond, VA, United States

Robert D. Nerenz, Department of Pathology and Laboratory Medicine, Dartmouth-Hitchcock Health System, Lebanon, NH, United States

James H. Nichols, Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, United States

Michelle Nieuwesteeg, Clinical Biochemistry, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada

Jason Y. Park, Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX, United States

M. Laura Parnás, Roche Diagnostics Corporation, Indianapolis, IN, United States

Khushbu Patel, Division of Pediatric Pathology, Department of Pathology, UT Southwestern Medical Center, Dallas, TX, United States

J. Peter R. Pelletier, Department of Pathology, University of Florida, Gainesville, FL, United States; Department of Anesthesia, University of Florida, Gainesville, FL, United States; Department of Pathology and Anesthesia, University of Florida, Gainesville, FL, United States

Morgan A. Pence, Clinical and Molecular Microbiology, Department of Pathology & Laboratory Medicine, Cook Children's Medical Center, Fort Worth, TX, United States

Lauren M. Petersen, Department of Pathology and Laboratory Medicine, Dartmouth-Hitchcock Medical Center and Norris Cotton Cancer Center, Lebanon, NH, United States; Geisel School of Medicine at Dartmouth, Hanover, NH, United States

Athena K. Petrides, Department of Pathology, Brigham and Women's Hospital, Boston, MA, United States; Harvard Medical School, Boston, MA, United States

Amy L. Pyle-Eilola, Department of Pathology and Laboratory Medicine, Nationwide Children's Hospital, Columbus, OH, United States

Alex J. Rai, Department of Pathology, Columbia University Medical Center, New York, NY, United States

Zahra Shajani-Yi, Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, United States

Lori J. Sokoll, Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD, United States

Joely A. Straseski, Department of Pathology, University of Utah School of Medicine, ARUP Laboratories Salt Lake City, UT, United States

Stefani N. Thomas, Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, MN, United States

Katie L. Thoren, Department of Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, United States

John Toffaletti, Duke University Medical Center, Durham, NC, United States

Steven M. Truscott, Department of Pathology and Laboratory Medicine, Oakland University William Beaumont School of Medicine, Beaumont Health, Royal Oak, MI, United States

Gregory J. Tsongalis, Department of Pathology and Laboratory Medicine. Dartmouth-Hitchcock Medical Center and Norris Cotton Cancer Center, Lebanon, NH, United States; Geisel School of Medicine at Dartmouth, Hanover, NH, United States

Alisha D. Ware, Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD, United States

Joesph R. Wiencek, Department of Pathology, University of Virginia School of Medicine, Charlottesville, VA, United States

Maria Alice V. Willrich, Department of Laboratory Medicine and Pathology, Division of Clinical Biochemistry and Immunology, Mayo Clinic, Rochester, MN, United States

William E. Winter, Department of Pathology, University of Florida, Gainesville, FL, United States; Department of Pediatrics, University of Florida, Gainesville, FL, United States; Department of Pathology and Pediatrics, University of Florida, Gainesville, FL, United States

Alison Woodworth, Department of Pathology and Laboratory Medicine, University of Kentucky Medical Center, Lexington, KY, United States

Alan H.B. Wu, Department of Laboratory Medicine, University of California, San Francisco, San Francisco General Hospital, San Francisco, CA, United States

Preface

It has been 4 years since the publication of the previous edition of the "White Book." As the field of clinical laboratory medicine continues to evolve, so has this book. With the release of the fourth edition of the *Contemporary Practice in Clinical Chemistry*, we have made some important changes. Notably, the book has moved from being published by our professional organization (AACC Press) to a major international publishing house (Elsevier), and we have added a Co-Editor to the book in Mark Marzinke. Both of these changes have been beneficial to the publication of this edition, and we are confident that these new collaborations will facilitate continued growth in future iterations of the "White Book."

Often times, when a major professional (or personal) task is accomplished, one tends to be satisfied with the result; however, it is also natural to think about improvement strategies or what could have been done differently. Our approach to this book has been no different. Consequently, in the fourth edition, we have tried to address any shortcomings or omissions from previous iterations, and we have augmented our exemplary roster of existing contributor authors with some of the newest and brightest in our field. We sincerely hope these changes have enhanced the content and communication of the material. This edition also includes new sections on Testing in Alternative Matrices, Applications of Mass Spectrometry, Clinical Hematology, and Clinical Microbiology.

This book is intended to be a supplement for the many other excellent training resources available in Clinical Chemistry. Therefore, the "White Book" aims to provide a clear and concise summary of a wide variety of topics to serve as a starting point for study and discussion. It is both our hope and goal that this resource will continue to be used by a wide variety of people in our field, from students to postdoctoral trainees to more experienced professionals and clinical laboratory scientists and directors.

A project like this does not happen without many experts willing to donate their time and share their knowledge on the written page—we are very grateful for each author who was willing to participate in the fourth edition of the "White Book." Thanks to you all! We also express our gratitude to our AACC colleagues and external subject matter experts within the field of Clinical Laboratory Medicine for the helpful feedback and suggestions, and for targeted criticism when needed. Last, we must thank our colleagues at Elsevier for their help in compiling all the chapters, reviewing of all the material, and keeping us moving toward completion—a transition between publishers is a big undertaking, and we appreciate their patience and support during the process.

March 2020

Bill Clarke and Mark Marzinke

Chapter 1

Preanalytical variation

Zahra Shajani-Yi and James H. Nichols

Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Understand why preanalytical variation is a significant contributor to laboratory errors.
- Identify the common sources of preanalytical variation.
- Discuss the potential effects of phlebotomy, tube additives, and order of draw.
- Discuss ways to detect and reduce preanalytical errors.

Laboratory testing comprises the majority of information in the electronic medical record [1,2]. Laboratory services accounted for 2.3% of health care expenditures in the United States with over 6.8 billion laboratory tests performed, with clinical pathology, anatomic pathology/cytology, and molecular/esoteric tests accounting for 66%, 23%, and 8% of performed tests respectively [3]. Hospital test volumes also grew by an average of 6% annually [3]. As the number of laboratory tests increases, the opportunity for errors that adversely affect patient care also increases. These errors can occur in any of the three phases of the total testing process: the preanalytic, analytic, or postanalytic phase (Fig. 1.1). An understanding of the sequence of events required for laboratory testing provides a foundation for assessing the likelihood of errors occurring at each step of the testing process.

This process begins as the clinician examines a patient and determines the need for a laboratory test. The correct test must be ordered, the patient must be prepared, and an appropriate sample must be collected. The sample is then transported to a laboratory, received, and processed for analysis. During analysis, the sample may be aliquoted, diluted, or subjected to subsequent testing before the final result can be verified for release. The clinician must then receive and interpret the result and decide on the appropriate treatment or follow-up and place the follow-up orders and instructions, and staff must schedule and carry out these orders for the patient.

Historically, quality initiatives have focused on the analytical phase of testing, and over the years, the number of errors attributed to this phase has decreased [4]. Interestingly, the majority of laboratory-related errors occur outside of the actual laboratory, either in the preanalytical or postanalytical phase. Recent studies report that approximately 46%–68% of all laboratory errors occur in the preanalytical phase [5,6].

Errors can occur during: (a) the ordering process, either through the clinician laboratory test order entry or when the order is manually transcribed; and (b) sample collection if a patient is not properly prepared or the sample is incorrectly labeled. Additional errors include: (c) specimen collection where specimens are either collected in the wrong type of tube with potentially interfering additives or if the tubes are collected in the wrong order; and (d) delays and/or inappropriate storage or handling during delivery of specimens to the laboratory. Finally, upon reaching the laboratory, testing accuracy is compromised if (e) samples are not adequately processed and stored for analysis.

Accurate laboratory test results demand high-quality specimens. Unfortunately, in most systems, the resources allocated for the pre- and postanalytical processes are not sufficient, as the importance of preanalytics is often overlooked. Many of the mistakes that are referred to as "laboratory errors" arise due to poor communication and action by others involved in the testing process or poorly designed processes that are outside of the laboratory's control. This chapter will focus on the most common sources of preanalytical variation and discuss some quality system processes for reducing the preanalytical errors. Understanding the causes of preanalytical errors coupled with proactive ongoing monitoring allows the laboratory to develop preventive measures to mitigate the risk of releasing inaccurate results.

Order entry

Errors in laboratory orders commonly occur due to the similarity of test names, improper use of synonyms,

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00001-6

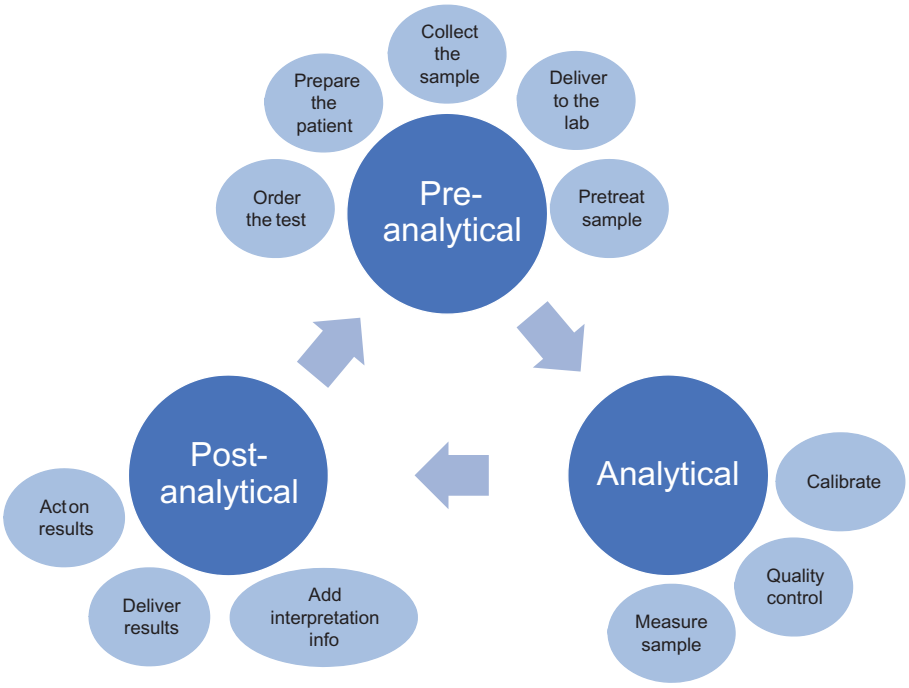


FIGURE 1.1 Steps in the testing process.

failure to enter orders correctly into the hospital electronic computer system, lack of knowledge about tests, and transcription errors (Table 1.1). Tests that are commonly misordered due to similar names are: (1) C-reactive protein for inflammation versus high-sensitivity C-reactive protein for cardiovascular risk assessment; (2) lipoprotein versus lipoprotein panel; (3) calculated versus direct low-density lipoprotein; and (4) 1,25-vitamin D (calcitriol) versus 25-vitamin D (calcidiol). Tests that also require supplementary clinical information often have high rates of errors. At our institution, in order to calculate a second trimester prenatal quad screen report, information such as the patient's date of birth, estimated due date, ethnicity, weight, diabetics, and smoking status must be provided by the clinician. Inaccurate reporting of such clinical data can lead to improper risk factors being calculated and subsequently reported.

Differences in methodology can also have implications for clinicians when they are trying to order a test. For example, testosterone can be measured accurately for most men by immunoassay, whereas women, children, and men with hypogonadism have lower testosterone concentrations and should therefore have testosterone measured by mass spectrometry. Ideally, a test name and description should be able to convey to the ordering provider if the test is appropriate for their patient. Further adding to these issues are the cases where orders are manually transcribed from written notes or requisitions, such as outpatient locations. These transcriptions are often performed in the specimen receiving section, where staff try to decode and/or determine what the clinician intended to order.

Redesigning the requisition and computer entry screens can greatly facilitate the order entry accuracy. In general, grouping common tests and liberal use of footnotes on written requisitions or pop-up screens for computer entry can enhance appropriate test selection by staff, but physicians begin to ignore these reminders due to information or alert fatigue. Expert systems and rules can check for duplicate orders and prevent common errors, like the addition of tests to an inappropriate tube type or collection of a specimen at an inappropriate time.

The key to improving the accuracy of test ordering is to implement a computer order entry system that requires physician ordering. This ensures that the correct physician and medical necessity are linked to each order for billing compliance. A physician order entry system also reduces the errors associated with verbal orders and manual transcription. Physicians often verbally dictate their test orders to residents and nursing staff who transcribe the tests onto written requisitions or computerized order entry systems. Staff may not be aware of the differences between specific tests or the effects of test methodology on results. Misunderstanding of a physician's request can also occur in the midst of a trauma or critical patient situation. A College of American Pathologists (CAP) Q-Probes study estimated that nearly 5% of physician requests were associated with one or more data entry errors [7]. At least 10% of the institutions surveyed had errors with one in five requisitions [7]. The study

TABLE 1.1 Sources of preanalytical variation.

Source of variation	Potential solution
Order entry	
Similar test names	Set up computer order entry screens with explanatory notes or pop-up screens
Duplicate orders	Construct expert systems and rules to detect duplicate orders
Transcription entry errors	Verify computer entry against written orders
Patient preparation	
Diet/supplements	Fast or restrict diet if necessary before testing Ask patients about supplements (e.g., biotin)
Time of collection	Proper collection for TDM and hormones Document drug administration accurately with respect to specimen collection
Specimen collection	
Patient identification	Verify the use of two identifiers
Needle size	Prevent hemolysis by routine smaller gauge needles
Tube selection/order of draw	Sign posted and smart laboratory labels
Prolonged use of tourniquet	Limit tourniquet use to 1 min
Fist clenching during phlebotomy	Encourage patients to rest arm during phlebotomy
Inadequate tube filling	Use vacuum collection tubes
Specimen clotting	Ensure tubes are more than 3/4 full during collection Mix tubes by gentle inversion immediately
Urine stability	Provide preservatives in collection container
Processing, transportation, and storage	
Outpatient clinic delayed processing	Provide equipment to process specimens on site
Exposure of tubes to environment	Protect and insulate specimens during transportation
Add-on testing	Validate and optimize storage stability for each analyte

concluded that verbal orders, high bed occupancy, and a failure to monitor the accuracy of order entry/having a policy to confirm correct order entry led to higher institutional error rates [7]. Accurate physician orders for laboratory testing are the first step in quality laboratory testing. Errors at the start of the laboratory test process can escalate to inappropriate specimen collection and the need to redraw a patient, leading to a delay in reporting test results and reaching a diagnosis, and subsequently delays implementing treatment.

Patient preparation

Once a correct order is placed, the patient must be prepared so that the results can be properly interpreted. Factors such as diet, exercise, medications, and time of collection (morning vs evening) can affect many laboratory tests. For some analytes, these effects are well-known—for example, ingestion of a meal will cause an increase in glucose and triglycerides, yielding values that are outside the fasting reference interval. For other analytes, the effects may not be recognized by physicians as significant contributors to result in variation, such as posture and exercise. A set of books by Donald Young is available that cites the basic scientific literature for general effects of drugs, disease, and preanalytical variables on clinical laboratory tests [8]. The manufacturer's package insert also describes the conditions and limitations for interpreting the result.

Catecholamines, epinephrine, and norepinephrine rise in response to smoking, exercise, stress, and ingestion of caffeine. In addition, cocoa is known to contain catechols that stimulate increases in catecholamines. Dietary restriction of chocolate- and caffeine-containing sodas, coffee, and tea is recommended for several days before the collection of 24-hour urine or plasma catecholamine specimens. Exercise can affect laboratory tests stimulating the release of catecholamines and other hormones, including β-endorphin, cortisol, glucagon, growth hormone (GH), and prolactin. Strenuous exercise works in muscles and can lead to increased levels of muscle enzymes, including creatinine kinase, creatinine, and aspartate aminotransferase (AST). Posture and moving from a recumbent to standing position cause shifts in fluid to the lower extremities and decrease plasma volume by ∼14% after 30 minutes. These fluid shifts concentrate on protein and lead to increases in serum osmolality, albumin, α_2-macroglobulin, transferrin, and total protein after 30 minutes of

standing. Protein-bound analytes, like calcium, also increase in proportion to the hemoconcentration produced by an upright position.

Timing of specimen collection is a further consideration and is particularly essential for therapeutic drug monitoring (TDM). Most therapeutic drug reference intervals or normal values are standardized to trough (predose) concentration for patients on an intermittent dosage regimen at steady state. The trough level is the lowest concentration of drug before the next dose. After consumption, drug levels will rise to a peak concentration and then slowly decrease as the drug is eliminated. The next dose repeats the rise and fall of drug concentration, as the drug is absorbed and eliminated. Collection of TDM specimens just before a dose will guarantee the trough concentration. Collection at other times will generate higher than expected levels that could be misinterpreted as too high or a possible overdose. Aminoglycoside antibiotics with concerns for both toxic side effects and minimal inhibitory concentrations to provide effective bacteriostatic and bactericidal activity may require both peak and trough levels for patient management. The peak concentration is generally considered to occur 30−60 minutes after completion of an intravenous (IV) dose of aminoglycoside antibiotics. Timing of peak specimens is therefore as critical as with trough specimens. Effective communication is required between the staff administering the medication and the phlebotomists collecting the specimen to ensure that an appropriate specimen is collected. Errors often occur when specimen collections are made with respect to when the drug is scheduled to be given rather than when the drug is actually administered. Good documentation, procedures, and communication can prevent such misunderstandings.

Most hormones have concentrations that vary throughout the day or month, and the reference ranges established for interpretation assume that the samples are collected at the proper time. Some hormones are released in a pulsatile manner and exhibit diurnal variation, with values that are higher in the morning and lower in the evening and overnight. In normal healthy patients, adrenocorticotropic hormone (ACTH) and cortisol concentrations are higher in the morning (between 04:00 and 12:00 hours) and have lower concentrations in the late evening and overnight. Iron also demonstrates diurnal variation with peaks between 08:00 and 16:00 hours and a nadir between 18:00 and 04:00 hours. Conversely, GH demonstrates a maximum peak 2 hours after the commencement of sleep. Follicle-stimulating hormone and luteinizing hormone have daytime peaks that are 15%−18% higher than nocturnal lows in women during the early follicular phase. These daily fluctuations are in addition to the wider variations seen during the follicular, midcycle, and luteal phases of the monthly menstrual cycle. Interpretation of the fertility hormone levels must be made with respect to the phase of the menstrual cycle and time of the day that the specimen is collected. Stress from illness and bed confinement can lead to increased cortisol, glucose, and insulin and decreased GH and thyroid-stimulating hormone. The collection of some laboratory tests, like glucose tolerance testing for the diagnosis of diabetes mellitus, should wait until the patient is discharged from the hospital. Patient preparation is an important consideration for guaranteeing a quality specimen and interpretable laboratory results.

Supplements can also affect laboratory results; in particular, biotin (also known as vitamin B7) supplement use has been increased, as biotin has recently been promoted for thickening hair, strengthening nails, and improving skin. Previously, high-dose biotin therapy was confined to patients with biotinidase deficiency, certain forms of alopecia, or multiple sclerosis. Many immunoassay manufacturers rely on biotin−streptavidin capture in their assays, and patients taking biotin may have a falsely increased result in competitive assays and a falsely low result in sandwich assays. The popularity of biotin supplementation has led to an increase in the reported cases of biotin interferences, increasing the potential for misdiagnoses [9]. Clinicians and patients should be aware that biotin can interfere with their test results and refrain from biotin supplements for approximately 1 week before obtaining laboratory testing on a platform that uses biotin−streptavidin capture. At the minimum, specimens should be collected at least 8 hours after the last dose of biotin, with the realization that residual biotin can affect some immunoassays.

Urine specimens can be affected by other factors. The concentration of analytes in urine is subject to renal function (ability of the body to eliminate the analyte) and the amount of water in the sample. Urine concentration is subject to hydration status, and urine is most concentrated in the morning and less concentrated as the subject is eating, drinking, and moving throughout the day. The concentration of an analyte like a drug may be less concentrated in the afternoon compared with the morning simply due to changes in urine water and the subject's hydration state. For this reason, many tests results are reported as ratios with respect to the concentration of creatinine that is eliminated relatively constantly based on the subject's renal function. Urine tests are also averaged over a 24-hour collection period to minimize further the effects of differences in a patient's hydration over the day. Collection of samples less than or greater than a 24-hour period of time can affect the daily average elimination estimates for urine tests.

Specimen collection

Once a physician order has been placed and the patient is appropriately prepared for the test, a proper specimen

must be collected at a suitable time. Patient identification is mandatory before any medical procedure including phlebotomy and laboratory testing. The Joint Commission has made patient identification one of its top patient safety goals, mandating the use of two unique identifiers to confirm positively the identity of a patient, such as the patient name in addition to the medical record number or another unique identifier [10]. Although it may appear to be easier to preprint labels before the time of collection, this practice is strongly discouraged. Preprinting labels will lead to samples with improper collection times and could lead to mislabeled samples, when one patient's label is applied to another patient's specimen. The former can lead to specimen rejection or result misinterpretation, whereas mislabeled samples can lead to a misdiagnosis. Given the severity of consequences from mixing up or mislabeling a laboratory specimen, the need to identify positively the patient cannot be overemphasized. Newer portable barcode systems are being deployed in hospitals, allowing the phlebotomist to scan a patient's wristband and print barcode labels at the bedside on demand. These systems are connected to the laboratory and hospital information systems and can stamp actual collection times and track specimens during transport and analysis in the laboratory.

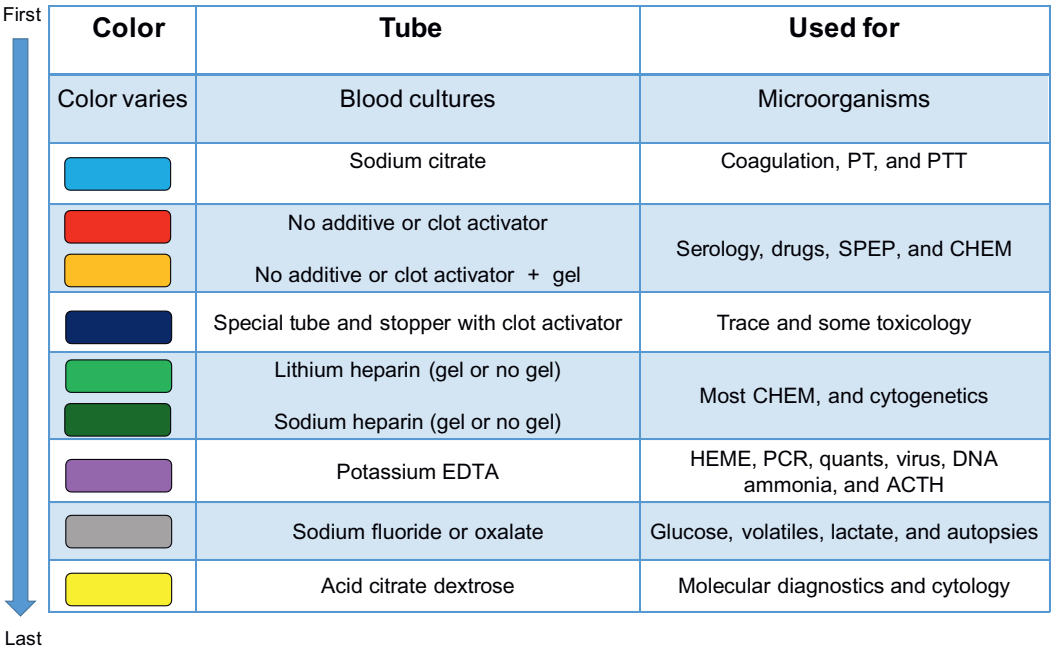
Tube type and order of draw

A number of tubes are available for specimen collection. Red-stopper tubes contain no additives or a clot activator, whereas other types of colored stoppers indicate the type of blood preservative or anticoagulant. Tubes are also available with a gel barrier that helps separate red blood cells from serum and plasma during centrifugation. Gel-barrier tubes provide the advantage of saving labor by allowing analysis directly from the collection tube. The primary tube can be centrifuged, analyzed, and stored without the need to aliquot the serum and plasma away from the cells. Chemistry analytes can be collected in red-top serum tubes, gold-top serum tubes (also known as serum-separator tubes), or green-top heparin preservative tubes for plasma analysis. Hematology specimens for cell counts require purple-top ethylenediaminetetraacetic acid (EDTA) preservative tubes, and coagulation specimens require blue-top citrate preservative tubes. Light-green, lithium heparin tubes are utilized for chemistry and electrolyte panels, while dark-green, sodium heparin tubes are utilized for drug analysis. The sodium content in the dark-green tubes will interfere with sodium analysis on chemistry and electrolyte panels of tests, while the lithium in the light-green tubes will interfere with lithium analysis. There are also special tubes for trace metals and toxicology (royal-blue top; manufactured to limit sample contamination with metals), blood culture (manufactured for sterility and microbiologic analysis), and glucose tolerance testing (gray top; manufactured with fluoride to inhibit glycolysis after sample collection). Drugs, proteins, lipids, and other analytes can bind to the gel barrier and may require red-top no-additive tubes without gel. Newer collection tubes contain plastic (ring-based) barriers and provide a physical separation of cells upon centrifugation, and use of plastic prevents drug and protein binding, unlike the gel-barrier tubes. A variety of different tube additives each pose specific test limitations, and this is one of the reasons why a universal anticoagulant tube has not been developed that is applicable to all tests. Whenever contemplating a change in the tube type or manufacturer, the laboratory should perform a thorough validation to determine the suitability of specimens received or stored in the laboratory before analysis.

The tube collection order used during phlebotomy can lead to further preanalytical errors. Phlebotomists are trained to collect blood culture and citrate tubes for coagulation measurements (light-blue top) first, followed by nonadditive serum (red-top) tubes. Heparin tubes (green), EDTA tubes (lavender), and other additive tubes are collected last (Fig. 1.2). Tubes should be collected in this order so that preservatives from the previous tube do not carry over to contaminate the next tube, as additive carryover can affect some laboratory tests [11]. For example, potassium EDTA binds divalent cations and alters coagulation results and analytes that require calcium and magnesium, such as potassium, calcium, and analytes such as AST, alanine aminotransferase, and alkaline phosphatase. Heparin can further carry over and impact coagulation testing if a heparin sample is collected prior to a sodium citrate tube. Correct order of draw can become a problem when blood draws are decentralized to nursing and clinical staff who may not be as familiar with the potential for laboratory interference from the tube-additive carryover.

Needle size, tourniquet use, and line collection

Appropriate selection of equipment is essential to the collection of a good specimen, specifically the size of the needle. Phlebotomy needles range from 18- to 23-gauge, with 25-gauge needles attached to butterfly-winged collection sets; smaller numbers indicate larger diameter needles. In general, the larger diameter 18-gauge needles are used for blood donor collection, and smaller 21- and 23-gauge needles are used for routine laboratory specimen collections. The smaller 25-gauge butterfly collection sets are reserved for difficult sticks from geriatric, cancer, and pediatric patients. Newer needles are manufactured with sheaths that cover or retract the needle after use to prevent accidental needlesticks. Needles also come in a variety of lengths depending on the depth of the vein to be punctured. Although smaller needles are less painful for the



Color	Tube	Used for
Color varies	Blood cultures	Microorganisms
	Sodium citrate	Coagulation, PT, and PTT
	No additive or clot activator No additive or clot activator + gel	Serology, drugs, SPEP, and CHEM
	Special tube and stopper with clot activator	Trace and some toxicology
	Lithium heparin (gel or no gel) Sodium heparin (gel or no gel)	Most CHEM, and cytogenetics
	Potassium EDTA	HEME, PCR, quants, virus, DNA ammonia, and ACTH
	Sodium fluoride or oxalate	Glucose, volatiles, lactate, and autopsies
	Acid citrate dextrose	Molecular diagnostics and cytology

FIGURE 1.2 Order of draw.

patient, they also increase the risk of sample hemolysis and the lysis of red blood cells in the sample, which could interfere with laboratory analysis. Smaller bore needles exert more shear stress on cells, leading to increased hemolysis. Phlebotomists must be alert during specimen collection, because small needles, difficult sticks, and a slow flow of blood can cause specimen hemolysis, which can lead to test interference depending on the analyte and may require recollection.

The prolonged use of a tourniquet can lead to hemoconcentration and pooling of blood. Use of a tourniquet for over 1–3 minutes can cause elevations in protein and albumin, calcium, potassium, and hemoglobin. It is recommended that phlebotomists should have everything in place prior to placing the tourniquet in order to minimize the time that the tourniquet is in place. Fist clenching during phlebotomy with a tourniquet in place can further lead to increases of 1.0–1.4 mmol/L in potassium levels.

Using collection tubes with a vacuum in the tube allows for the appropriate amount of blood to be collected, and manufacturers have designed collection systems to facilitate the collection of multiple tubes. Use of a syringe to collect blood that is then injected through the stoppers of multiple tubes can lead to specimen hemolysis, filling errors, cross-contamination, and loose stoppers that can leak during transportation to the laboratory. The use of a needle and a syringe to puncture the stopper should be strongly discouraged, as this practice is dangerous and poses a risk of needlestick injury. In addition, tubes containing additives must be thoroughly mixed after collection for the additive to be distributed; undermixing can lead to specimen clotting, whereas overmixing can induce hemolysis. Phlebotomists are encouraged to invert gently each tube 5–10 times after collection.

Additional problems can be encountered when collecting specimens through lines or catheters. The presence of an IV, arterial, or catheter line does offer staff easy access to the patient, but the practice of collecting specimens through lines and catheters should also be strongly discouraged due to the potential for contamination and dilution [12]. IV lines contain fluids that are being given to the patient and pose a risk of diluting a specimen that is being collected through a line or increasing the concentration if the analyte to be tested is contained in the IV fluid, such as sodium or glucose. Lines are meant to administer fluids to a patient and are not intended for specimen collection. Some newer catheters are made of a plastic material that is rigid for insertion but becomes pliable when the catheter reaches body temperature. This type of line is particularly problematic, because it is directional. Although fluids can be administered and the line stays open, drawing backward causes the line to collapse, making specimen collection difficult and increasing the risk of hemolysis. Heparin is used to flush and keep the access open on some catheters, particularly heparin locks with arterial access lines; however, heparin can interfere with other laboratory tests. It is understandable why clinical staff are hesitant to stick a patient with a line already inserted; however, discarding a small amount of blood as a line flush before collecting a specimen should be discouraged, as it does not guarantee that a proper sample

can be obtained. D25W (25 g of dextrose per 100 mL of solution), vancomycin, and several other large drugs and antibiotics tend to stick to lines and cannot be adequately flushed. Carryover of line fluid requires the laboratory to cancel the specimen and request for a recollection of the test, which both inconveniences the patient and delays the turnaround time for test results [13].

Staff training and periodic competency evaluation is suggested to ensure consistent specimen quality from all staff involved in specimen collection. Unfortunately, even with standardized training and periodic competency assessments, staff can continue to make errors during specimen collection. Our laboratory, for instance, has experienced several cases where staff collected one tube and, after phlebotomy, realized that they needed a different tube, resulting in one tube being poured into a tube with another additive. In another instance, a microtainer was inadvertently collected in a purple-top EDTA additive tube. When the staff realized they needed a gel-barrier tube, they opened the cap and inserted the purple-top EDTA microtainer into a gel-barrier serum tube and sent it to the laboratory. These staff did not realize that the tubes contained different additives that could affect the test results (Fig. 1.3). We even had two recent cases where a specimen was received and centrifuged, and the red blood cells were delivered to the technologist instead of plasma. Even with adequate training, staff should have regular retraining and constant reminders of the proper procedure to follow.

There are some technologies available to ensure staff collect the correct tube from the correct patient in the proper order and thus reduce phlebotomy-associated errors. Several companies now distribute portable personal data-assistant devices that can upload a list of patients requiring phlebotomy. The devices contain a barcode scanner, so staff can scan their identifications to access the software and access the lists of patients to collect. Each phlebotomist's identification is then linked to the specimen collection event and tubes collected at that time. This provides a tracking mechanism required for laboratory accreditation and troubleshooting phlebotomy errors. After staff positively identifies the patient with two unique identifiers, the device can scan a patient wristband, and the patient's name will appear (as additional verification of identification) with a list of laboratory tests pending, tubes to be collected, and proper order of draw. As each tube is collected, staff can scan the tube and a label is printed at the bedside, reducing the possibility of mislabeling and capturing the exact time of specimen collection. After the tubes for a patient have been collected, the device can be downloaded to a laboratory information system, which captures the collection information and acts as a tracking mechanism for the transportation of the specimens to the laboratory. Manufacturers even offer wireless versions of these collection devices. Implementation of these systems has reduced phlebotomy and labeling errors and has the potential to reduce specimen recollections and delays in result turnaround time.

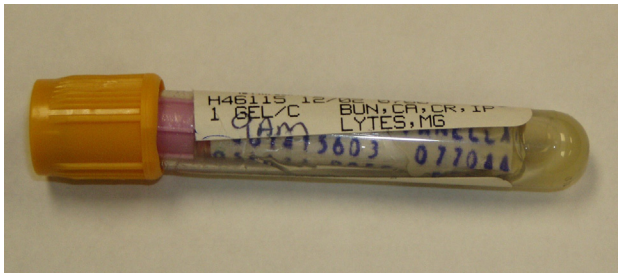


FIGURE 1.3 Inadvertent specimen collection: a purple-top EDTA microcontainer was collected from a patient. After phlebotomy, staff realized they needed to collect a gel-barrier tube instead; therefore they pushed the purple-top microcontainer into the large gel-barrier tube and sent it to the laboratory. The staff did not understand when they were called to cancel the specimen, claiming that the specimen was in a gel tube.

Processing and transportation

After collection, specimens should be transported to a laboratory, processed, and analyzed as soon as possible as delays can compromise the results. Glucose, for instance, decreases at a rate of 5%−7% per hour in whole blood at room temperature [14]. Glycolysis will continue until the specimen is processed by centrifugation, and serum and plasma are separated from the cellular components of blood. Patients with increased white blood cell counts have a higher number of cells and greater metabolic demand, so glycolysis in their whole blood will be faster than that in specimens from normal patients.

Other analytes are unstable in unprocessed blood or when stored at room temperature instead of refrigerator or freezer temperatures. Specimens need to be processed by centrifugation as soon as possible after collection (for plasma) or specimen clotting (for serum). For most chemistry analytes, laboratories should seek to process blood by centrifugation within 30−60 minutes of collection, with serum and plasma aliquoted (for tubes without gel barriers) and stored refrigerated until analysis. Ammonia, however, is an unstable analyte, and increases beyond the total allowable error limit can be seen after just 30 minutes at room temperature. Samples are recommended to be chilled, transported on ice, centrifuged within 15 minutes of collection, aliquoted off the cells, and plasma analyzed within 90 minutes to 2 hours when stored refrigerated. Care must be taken to centrifuge specimens only once, as recentrifugation can release cellular components like potassium and lactate dehydrogenase, and the plasma sample should remain undisturbed and vertical after centrifugation. Remixing plasma gel samples after

centrifugation has been recently shown to cause falsely increased values of 25-vitamin D on some assays due to resuspension of cells and platelets [15]. For this reason, a proper technique when processing plasma samples is essential; all samples should be aliquoted and not poured over so that cell debris and particulate matter do not enter the sample and compromise the results. This can be an issue with pediatric samples collected in microcontainers, where well-meaning staff members are so concerned that a sample will be canceled for insufficient volume that they will invert the tube and tap to ensure that every last drop is aliquoted, carrying debris and particulates that float above the separator after centrifugation. This cellular debris can interfere with chemistries as well as immunoassays on automated instrumentation.

Many hospitals rely on automated processes to transport samples to the laboratory, such as pneumatic tube systems that connect the laboratory processing areas directly to the inpatient medical units. For outpatient testing, clinics and phlebotomy stations should be provided with centrifuges and equipment to process the specimens on site before transportation to the core laboratory by courier in order to minimize sample instability. For example, if glucose is not appropriately processed at an outpatient collection facility, glucose values will continue to decrease and the result reported may be (falsely) critically low.

There are other laboratory tests that require special handling, processing, and storage. Hematology testing is performed on whole blood samples and thus these do not require centrifugation, but must be well mixed before analysis by automated cell counting systems. Hematology samples also must be protected from freezing that could result in hemolysis. Samples for metanephrines, on the other hand, require prompt centrifugation and freezing of serum and plasma to maintain stability. Additional attention may be needed to protect samples for bilirubin analysis from light during transportation, processing, and storage. Ammonia and hormones such as ACTH, gastrin, and calcitonin are unstable and must either be tested or frozen immediately. For urine specimens, there are a variety of preservatives available, depending on the stability of the analyte. Some urine preservatives are acceptable to add to the sample within 2 hours of the completion of a 24-hour or random urine collection, whereas other preservatives must be added before sample collection to prevent microorganism growth, prevent ion precipitation, or to preserve the sample for urine culture. The optimal preservative will depend on the laboratory's analytical methodology and recommendations for one laboratory may not be acceptable to another laboratory due to differences in analysis.

Special consideration and equipment may need to be made for collection, processing, and transportation of laboratory samples at outpatient locations. Clinics may need to have centrifuges, freezer capabilities, refrigerators, and storage space for preservatives, collection containers, and phlebotomy supplies. In the summer, samples may need to be transported in coolers with ice packs to maintain constant temperature and protect from overheating, whereas in the winter, samples may need to be transported in coolers (without ice or cold packs) to protect samples from freezing temperatures. Delays in processing or transportation can also further compromise the specimen and analyte stability [16].

Laboratories must also arrange for appropriate storage of specimens after analysis. Clinicians frequently add - on tests after the initial orders have been completed. Most chemistry samples can be stored for 7 days when refrigerated, allowing for specimen retrieval when additional testing is ordered; however, adequate storage space is necessary. Stored specimens are also useful to troubleshoot questionable results and for legal documentation, as the samples can be retrieved and analysis repeated when necessary.

Detecting preanalytical errors

Although phlebotomist training, prompt processing, and transportation to the laboratory are the first line of defense for preventing preanalytical errors, the laboratory must also create quality systems to inspect routinely specimens for common preanalytical errors. One obvious error is collecting a specimen in the wrong type of tube. If a potassium ethylenediaminetetraacetic acid (K_2EDTA) (purple-top) hematology specimen is sent to chemistry for an electrolyte or potassium order, the laboratory can recognize that the wrong type of tube was collected and can contact the ordering physician, cancel the specimen, and request a recollection in the correct tube type. An issue that cannot be detected prior to analysis is the scenario where a clinic mistakenly collects the K_2EDTA specimen, processes the specimen, and sends a plasma aliquot for electrolyte analysis but labels it as a serum specimen. This type of error will be detected only after analysis, and only if upon review of the results, a technologist realizes that the potassium level has increased into the critical and nonphysiologic range. Fortunately, laboratory information system rules can also be set up to flag erroneous results, such as test results that are nonphysiologic or unable to support life, results that are in a life-threatening critical range either high or low (critical error flags), or results that differ significantly from previous results (delta checks). Erroneous and critical result flags should call attention to potential preanalytical issues with sample contamination from inappropriate tube additives, dilution with IV fluid, and specimen clots or other sources of preanalytic error. The delta check flag calls attention to test

results that differ from a previous result by more than a predetermined limit. Delta checks help in the detection of preanalytical errors, such as mislabeled specimens, but are limited to patients who have a history of previous results for comparison.

The presence of specimen interference is another common source of error. Some interferences may be metabolic (increased urea) or disease-related (increased bilirubin), whereas other interferences may be the consequence of failing to prepare the patient (lipemia and nonfasting samples), a difficult phlebotomy (hemolyzed red blood cells), or administration of certain therapies (drugs or monoclonal antibodies) or supplements (vitamin B7 or biotin). Serum indices are a spectrophotometric estimate of specimen interference from icterus (presence of bilirubin), lipemia (lipids and chylomicrons), or hemolysis (hemoglobin in serum). The presence of significant interference used to be handled by the visual inspection of the sample, and technologists traditionally handled each specimen and appended a note to those results that might be affected by apparent interference, yet this method is highly subjective. Today, although fewer specimens are handled directly, a quantitative estimate of icterus, lipemia, and hemolysis can be measured with every patient specimen on high-volume automated chemistry analyzers. This measurement is based on the ratio of sample absorbance at various wavelengths multiplied by a correction factor (Fig. 1.4). Manufacturers publish acceptance thresholds for icterus, lipemia, and hemolysis that will cause clinically significant interference for tests in the method's package insert; however, it is strongly recommended that laboratories verify the manufacturer's recommendations by constructing their own interference curves or matrices.

Serum interference curves can be constructed by performing an experiment outlined in the Clinical Laboratory Standards Institute (CLSI) EP07 protocol for dose response testing [17]. A series of test samples, systematically varying only in the concentration of interferent, is prepared by making quantitative mixtures of two pools, one at the highest concentration of the interferent to be tested, and the other at the lowest [17]. Alternatively, varying amounts of interferent can be spiked into a sample pool to create a set of samples with the same analyte concentration but varying amounts of interferent. Creation of a sample set requires that each sample be spiked with the same volume of the interferent, so there is no dilutional difference between the individual samples, and the total volume added is only a small proportion (ideally $<10\%$) of the sample so that the sample matrix is not significantly altered. All samples are analyzed in one analytical run, and the results are graphed to determine the level of interferent required to cause significant bias or shifts in the true test results. If there is no interference, all samples will generate the same test result within assay precision. Interference is characterized by a difference in results from baseline (zero interferent sample) with samples containing increasingly higher concentrations of interferent. Interferences can be positively or negatively biased, generating higher or lower results in proportion to the amount of interferent (Fig. 1.5). Tables can then be constructed to summarize the interference cutoffs or limits

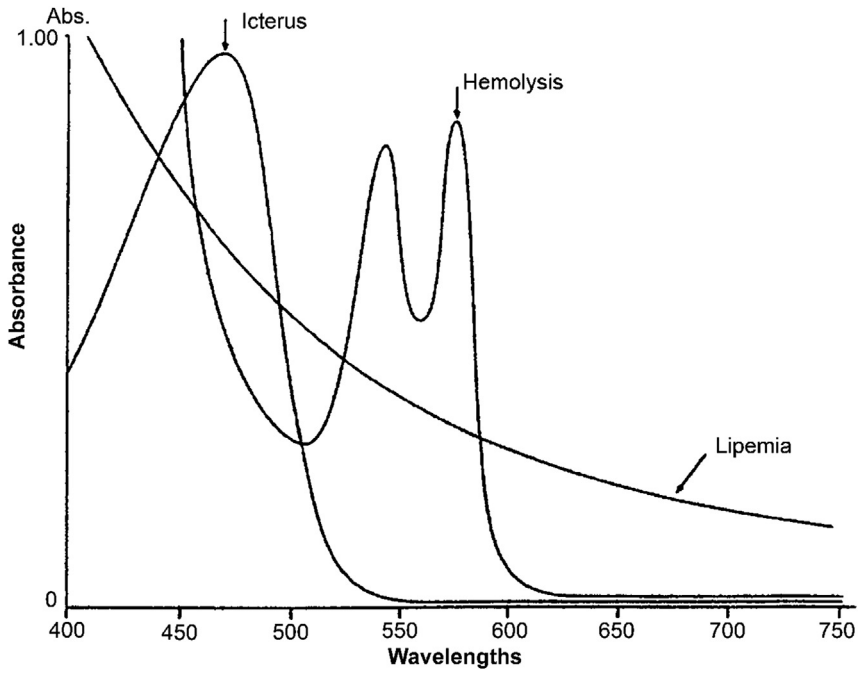


FIGURE 1.4 Serum indices. Interference in a serum or plasma specimen can be determined spectrophotometrically by a ratio of wavelengths and a correction factor that links absorbance to the intensity of interference. *Reproduced with permission from Clinical Chemistry Learning Guide Series, Abbott Diagnostics (2017).*

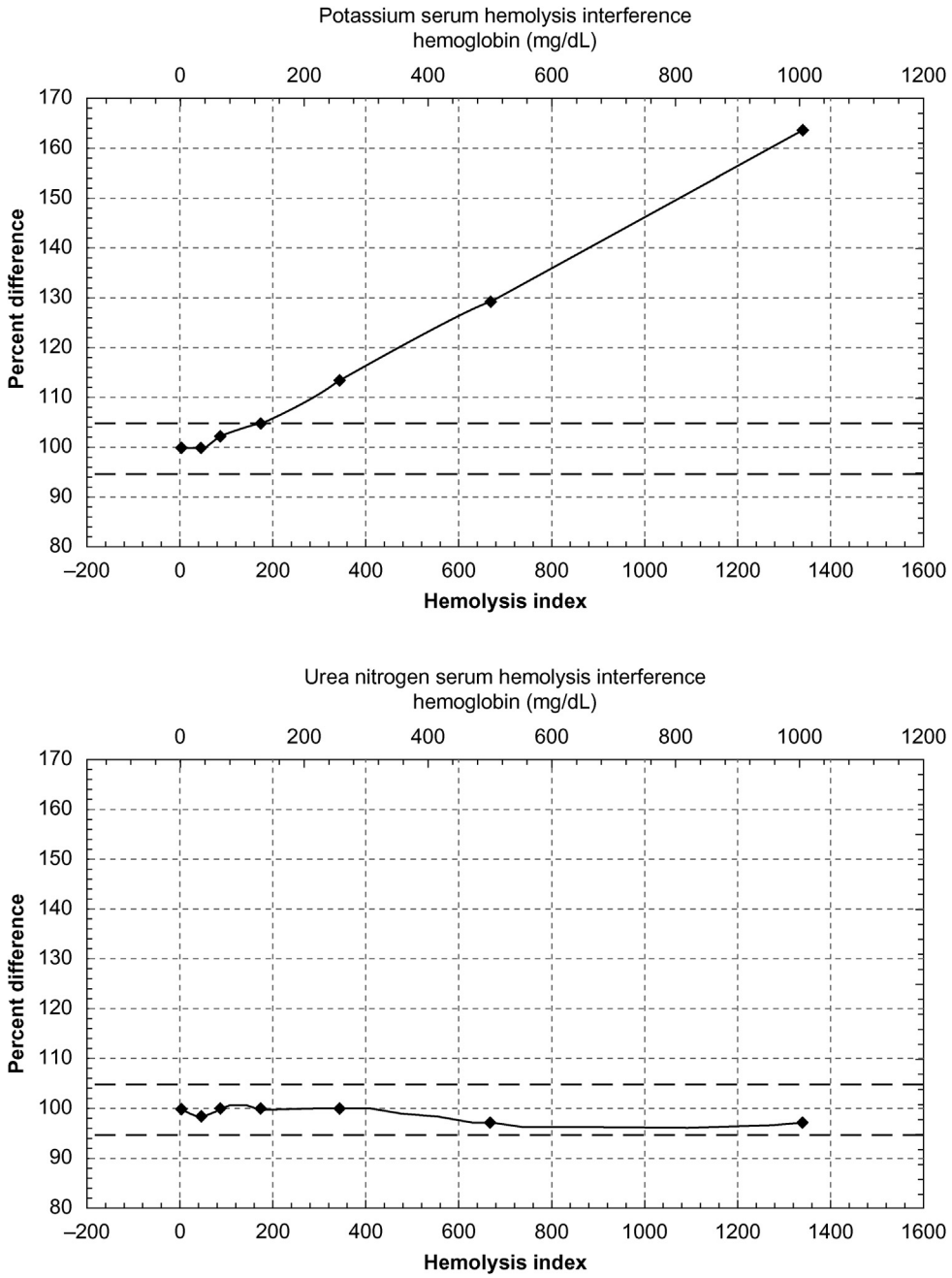


FIGURE 1.5 Serum interference curves. Hemolysis interference curves for potassium and urea nitrogen in serum. Hemoglobin (mg/dL) *(upper x-axis)* is the amount of hemoglobin spiked into a serum sample pool, whereas the hemolysis index *(lower x-axis)* is a spectrophotometric estimation of interference level (color intensity) from an automated chemistry analyzer. Dashed lines indicate our laboratory's tolerance limits for clinical significance as a percentage of difference from a sample with no interferent (e.g., 5% for potassium and urea nitrogen).

for each analyte (Table 1.2). For patient specimens, the serum indices can be compared with the predefined interference limits to determine whether the level of interference is clinically significant. The ability to conduct quantitative serum indices on every sample has removed the historical subjectivity from technologists visually estimating specimen quality. Serum indices add to critical values, delta checks, and other flags that can detect potential specimen issues.

Unfortunately, there are few ways to detect reliably the metabolic or drug interferences unless the interferent is colored or the interference generates an error during analysis. Sometimes, these errors can be detected by delta checks if the drug or the metabolite changed, since a recent test was conducted and the interference is significant. In other instances, a clinician may call the laboratory and indicate that the test result does not match the patient's clinical symptoms. Establishing and maintaining a good relationship with clinicians are important factors to establishing a quality laboratory service. The laboratory relies on clinicians to raise issues and provide feedback and is best served when clinicians question unusual

TABLE 1.2 Serum interference tables.

Icterus interference					
Test	**Limit I index**	**Laboratory tolerance (%)**	**Direction**	**Manufacturer index limit**	**CAP limit (%)**
Urea	29	<5	↓	60	± 9%
Potassium	53	<5	↔	Not significant	± 0.5
Cholesterol	15	<3	↓	25	± 10%
Glucose	15	<5	↓	60	± 10%
Lipemia interference					
Test	**Limit L index**	**Laboratory tolerance (%)**	**Direction**	**Manufacturer index limit**	**CAP limit (%)**
Urea	1119	<5	↔	1000	± 9%
Potassium	1119	<5	↔	Not significant	± 0.5
Cholesterol	1119	<3	↔	1250	± 10%
Glucose	1119	<5	↔	1000	± 10%
Hemolysis interference					
Test	**Limit H index**	**Laboratory tolerance (%)**	**Direction**	**Manufacturer index limit**	**CAP limit (%)**
Urea	1337	<5	↔	1000	± 9%
Potassium	175	< 5	↑	At all levels	± 0.5
Cholesterol	670	<3	↔	700	± 10%
Glucose	670	<5	↓	1000	± 10%

Index limits (icterus, lipemia, and hemolysis) for the selected analytes indicate our laboratory's tolerance for clinical significance (percentage of bias from a sample with no interferent). Manufacturer limits (from package insert) and CAP proficiency survey tolerance recommendations are shown for comparison. Interferences noted are only an example and not a representative of a specific manufacturer or instrument model. Direction of interference is included: ↔, no change; ↑, positive interference and high bias; ↓, negative interference and low bias. *CAP*, College of American Pathologist.

results, discuss them with the laboratory, and work together to resolve the source of the problem. Manufacturer's package inserts are vital sources of information when results are questioned, and most package inserts indicate the levels of drugs, metabolites, icterus, lipemia, or hemolysis that can cause significant interference. Stored specimens are an additional resource when investigating potential interferences, since stored specimens can be retested by the same methodology or sent to another laboratory for analysis by a different methodology if necessary.

Summary

Preanalytical variation is an important source of laboratory errors. The laboratory is responsible not only for the analysis of a specimen but also for ensuring control over preanalytical, analytical, and postanalytical processes to guarantee quality results for patient care. Understanding the sources of preanalytical variation and taking steps to minimize potential for errors before the specimen arrives in the laboratory are efforts that can reduce the need for specimen recollections, minimize delays in the turnaround time of test results, eliminate unnecessary medical follow-ups, and facilitate improved patient outcomes. Some preanalytical interferences such as icterus, lipemia, or hemolysis can be detected by the laboratory, preventing results from being released and warning clinicians of the potential for interference; unfortunately, other errors and interferences may not be easily detected. Too often, not enough attention is placed on minimizing the causes of preanalytical errors. Preanalytical errors can be decreased by having trained phlebotomists conduct blood draws, assuring that samples are transported to the laboratory in a timely fashion, having a well-trained specimen processing and receiving staff, and by maintaining a close working relationship between the laboratory and the information technology staff. The laboratory is an integral part of the health care team, and teamwork is required to ensure the integrity of specimen results from patient order through specimen collection, analysis, interpretation, and implementation of patient treatment. Preanalytical variation

is a significant component of the laboratory testing process and everyone, from clinicians to medical directors to nurses to laboratory staff, must be aware of the potential for error and work together to guarantee quality results. As the volume and complexity of testing increases, the laboratory and clinicians should foster a solid partnership to ensure appropriate test selection, interpretation, and use of the laboratory.

References

[1] R.W. Forsman, Why is the laboratory an afterthought for managed care organizations? Clin. Chem. 42 (1996) 813–816.

[2] M.J. Hallworth, The '70% claim': what is the evidence base? Ann. Clin. Biochem. 48 (2011) 487–488.

[3] The Lewin Group. Laboratory medicine: a national status report 2008. <http://http://www.lewin.com/content/dam/Lewin/Resources/Site_Sections/Publications/3993.pdf>, 2008.

[4] M. Plebani, The detection and prevention of errors in laboratory medicine, Ann. Clin. Biochem. 47 (2010) 101–110.

[5] G. Lippi, J.J. Chance, S. Church, P. Dazzi, R. Fontana, D. Giavarina, et al., Preanalytical quality improvement: from dream to reality, Clin. Chem. Lab. Med. 49 (2011) 1113–1126.

[6] M. Plebani, Errors in clinical laboratories or errors in laboratory medicine? Clin. Chem. Lab. Med. 44 (2006) 750–759.

[7] P. Valenstein, F. Meier, Outpatient order accuracy. A College of American Pathologists Q-Probes study of requisition order entry accuracy in 660 institutions, Arch. Pathol. Lab. Med. 123 (1999) 1145–1150.

[8] D.S. Young, Effects of Preanalytical Variables on Clinical Laboratory Tests, 2nd ed., AACC Press, Washington, DC, 2007, p. 1917.

[9] B.M. Katzman, A.J. Lueke, L.J. Donato, A.S. Jaffe, N.A. Baumann, Prevalence of biotin supplement usage in outpatients and plasma biotin concentrations in patients presenting to the emergency department, Clin. Biochem. 60 (2018) 11–16.

[10] Joint Commission Laboratory Services, National Patient Safety Goals, Joint Commission Resources, Oakbrook Terrace, IL, 2019. Available from: https://www.jointcommission.org/accreditation/lab_standards_information.aspx.

[11] Clinical and Laboratory Standards Institute, GP41 Collection of Diagnostic Venous Blood Specimens, 7th ed., CLSI, Wayne, PA, 2017.

[12] Nichols J.H., Rajadhyaksha A., Camelo-Piragua S., Rauch C.A. Indirect phlebotomy. Check Sample Clinical Chemistry No. CC 08-3 (CC-369) ASCP (2018).

[13] T.P. Darcy, S.P. Barasch, R.J. Souers, P.L. Perrotta, Test cancellation: a College of American Pathologists Q-Probes Study, Arch. Pathol. Lab. Med. 140 (2016) 125–129.

[14] N. Rifai, A.R. Horvath, C.T. Witter, Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, sixth ed., Elsevier, St Louis, MO, 2018.

[15] J.A. Hernandez, J.E. Stanford, C.V. Savoie, J.H. Nichols, J.M. Colby, Spuriously elevated 25-hydroxyvitamin D in lithium heparin plasma samples transported by courier, J. Appl. Lab. Med. 02 (2018) 807–813.

[16] J.R. Wiencek, J.H. Nichols, Impact of ambient seasonal temperatures on specimens stored in courier lockboxes, J. Appl. Lab. Med. (2018) 970–976.

[17] Clinical Laboratory Standards Institute, EP07: Interference Testing in Clinical Chemistry, CLSI, Wayne, PA, 2018.

Self-assessment questions

1. Which of the following contribute to preanalytical variation?
   - **a.** biologic variation
   - **b.** sample collection errors
   - **c.** delays in processing
   - **d.** all of the above
2. When do most errors in laboratory testing occur?
   - **a.** while ordering the correct test
   - **b.** before the sample arrives in the lab
   - **c.** during specimen analysis
   - **d.** during physician interpretation
3. CAP estimates what percentage of physician requests are associated with errors in data entry?
   - **a.** 5%
   - **b.** 10%
   - **c.** 20%
   - **d.** 25%
4. What can be done to enhance staff efficiency and ensure specimens are collected at the right time?
   - **a.** Preprint sufficient labels to allow staff to collect multiple patients during a phlebotomy round.
   - **b.** Use coat pockets to store supplies and collected tubes between the patients.
   - **c.** Communicate with staff and check medical record documentation prior to collecting specimens.
   - **d.** Ensure specimens are collected only when it is convenient for the patient.
5. Which of the following are possible resources for information on laboratory interference?
   - **a.** *Physician's Desk Reference* with drug inserts on prescription medications
   - **b.** Dr. Young's books on laboratory test effects
   - **c.** manufacturers
   - **d.** all of the above
6. Which of the following are possible sources of preanalytical variation?
   - **a.** season of the year
   - **b.** working in the night shift
   - **c.** sunlight
   - **d.** all of the above
7. Prolonged use of the tourniquet can lead to increases in which analyte?
   - **a.** calcium
   - **b.** digoxin
   - **c.** phosphate
   - **d.** triglycerides
8. Which of the following is most a concern for preanalytical variation?
   - **a.** collecting a specimen through an IV line
   - **b.** inadequate filling of the collection tube
   - **c.** patient identification
   - **d.** all of the above
9. Why are newer PDA technologies for phlebotomy collection an advantage?
   - **a.** to save time
   - **b.** to provide positive patient identification
   - **c.** to track the phlebotomist and specimen
   - **d.** all of the above
10. Which of the following tubes will clot?
    - **a.** green-top heparin
    - **b.** lavender-top EDTA
    - **c.** red-top activator
    - **d.** gray-top fluoride/oxalate
    - **e.** light-blue top citrate
11. What is hemolysis?
    - **a.** breakdown of red blood cells
    - **b.** a yellow by-product of hemoglobin breakdown
    - **c.** high triglycerides that make the sample turbid
    - **d.** an assay interference from radiopaque dye
    - **e.** a and b

Answers

1. d
2. b
3. a
4. c
5. d
6. d
7. a
8. d
9. d
10. d
11. a

Chapter 2

Statistical methods in laboratory medicine

Daniel T. Holmes

Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada

Learning objectives

After reviewing this chapter, the reader should be able to:

- Understand and apply descriptive statistical analyses and visualizations to a laboratory medicine data set.
- Describe how to determine if a data set is significantly different from the normal distribution.
- Understand the application of statistical tests of central tendency.
- Explain the differences between different linear regression methodologies.
- Discuss how to numerically evaluate diagnostic test performance.

Introduction

The purpose of this chapter is to introduce the major statistical concepts and tools required for clinical practice. This is not a resource as used in undergraduate statistics courses. There are no derivations or proofs and few explanations on how to perform statistical calculations by hand. There are many free resources and textbooks that do an excellent job of this. Rather, this chapter is intended to highlight important content and key conceptual notions. However, in the modern era, it is impossible to divorce statistical analysis from a software tool, and it is therefore necessary to pick a representative tool to accompany this chapter. Traditionally, this would have been the pervasively installed spreadsheet software. However, spreadsheet programs do not have the specialized statistical tools required for clinical practice. For this reason, I have elected to use the R statistical programming language to perform the calculations demonstrated. This does not mean that this chapter will introduce the R language, as there are many online texts [1,2] and open-learning resources [3] for this. Rather, by showing example R code, the student will be able to see that, in a few lines, a great deal can be accomplished and it is my hope to inspire readers to learn R for use in their clinical practice.

Basic descriptive statistical analysis

Consider a data set taken from the Intersalt study [4], where diastolic blood pressure (BP) and 24-hour urine sodium excretion were collected on individuals from several countries.

The data set happens to have 52 entries, but only the first 10 are displayed in Table 2.1.

Central tendency and dispersion

In lab medicine, we need to be able to calculate the mean, standard deviation (SD), median, interquartile range (IQR), and selected quantiles (usually expressed as percentiles).

TABLE 2.1 Representative data from the Intersalt study.

Diastolic pressure (mm Hg)	Sodium excretion (mmol/day)	Country
72.0	149.3	Argentina
78.2	133.0	Belgium
73.9	142.6	Belgium
61.7	5.8	Brazil
61.4	0.2	Brazil
73.4	148.9	Canada
79.2	184.3	Canada
66.6	194.1	Colombia
82.1	135.6	Denmark
75.0	138.7	East Germany

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00002-8

Mean

The mean of x, denoted $\bar{x}$, is calculated by:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.1}$$

that is, "add up all values of x and divide by the total number of values." The problem with expressing results as a mean is that the mean is strongly affected by outlier results. By way of example, turnaround times for lab tests should not be expressed by using the mean, because there are often problematic individual samples with very large turnaround times.

In R, the mean of the BP results can be calculated by reading the data, and by using the `mean()` function:

```
# Any text preceded by a "hastag" (#) is ignored by R and serves
# an informational comment.
library(DAAG) #load the package that contains the dataset
data(intersalt) #load the intersalt data
mean(intersalt$bp) #calculate the average blood pressure
```

```
## [1] 73.15192
```

The R output, shown here (and throughout this chapter) preceded by ## marks, shows us that the mean diastolic BP is 73.2 mmHg.

Median

The median is the middle value of the data set. That is, if we order the whole data set from lowest to highest values, the middle value, calculated as the $\left(\frac{n+1}{2}\right)^{\text{th}}$ entry, is the median. However, we encounter a problem when there are an even number of data points, because there is no middle value. In this circumstance, the median is given by the average of the two middlemost values. Our Intersalt data set has 52 values, so, if we sort the BP values and calculate $\left(\frac{n+1}{2}\right)$, we get the number 26.5, which is to say that the median lies halfway between the 26th and 27th values.

We can do this out the long way:

```
sort(intersalt$bp) #display the raw data sorted
```

```
##  [1] 61.4 61.7 62.9 66.1 66.6 67.2 67.4 67.9 68.0 68.4 69.9 70.0 70.2 70.7 71.2
## [16] 71.4 71.7 72.0 72.1 72.4 72.4 72.6 72.7 72.9 73.1 73.1 73.2 73.4 73.5 73.9
## [31] 73.9 74.7 75.0 75.2 75.5 75.6 75.7 76.0 76.2 77.2 77.4 77.5 77.9 78.2 78.2
## [46] 78.5 79.2 79.2 79.6 79.7 81.4 82.1
```

From the data, we can see that the 26th and 27th data points are 73.1 and 73.2, respectively. Therefore the median is 73.15. In the R language, this can be accomplished with the `median()` function:

```
median(intersalt$bp)
```

```
## [1] 73.15
```

Standard deviation, interquartile range, and quantiles

The SD is the square root of the mean square deviation of the results and is a measure of the dispersion of the data. Readers are no doubt aware that the SD has predictable meaning in normally distributed data sets (a concept to be discussed). The SD is defined as:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \tag{2.2}$$

where $\bar{x}$ is the mean value of x. Most textbooks quote the formula with $n-1$ in the denominator, which makes the result an unbiased estimate of the SD, accounting for limitations associated with finite data sets. It is therefore preferential to use the following:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}. \tag{2.3}$$

In the R language, the SD function is `sd()`, and the calculation is achieved as follows:

```
sd(intersalt$bp)
```

```
## [1] 4.786562
```

Quantiles and the interquartile range

Quantiles are a way of expressing how the data are distributed. The quantiles of the variable x are any set of intervals of x that define bins with an equal number of counts. The reader is likely familiar with the concept of percentiles, bins that each contain 1% of the data. Another common approach is the quartile, where there are four intervals of x to define bins containing 25% of the data.

Because the SD does not provide predictably interpretable information for non-Gaussian distributions, when the distribution has skewness or excess kurtosis, the IQR is frequently provided as an alternative. The IQR is the difference between the 75th percentile and the 25th percentile, thereby defining a range of x encompassing the central 50% of the distribution, not necessarily centered about the mean and median of the distribution.

In the R language, the IQR is determined by using the `IQR()` function:

```
IQR(intersalt$bp)
```

```
## [1] 5.875
```

However, if there is a desire to calculate a specific quantile, this can be achieved with the `quantile()` function. There are a number of strategies for inferring the quantiles of a distribution, and R implements nine of them, any one of which can be specified if desired. The simple strategy of linear interpolation, recommended by the Clinical Laboratory Standards Institute (CLSI) Document EP28-A3C [5], is available, but not as default.

The following R code calculates estimates of the 10th, 25th, 50th, 75th, and 90th percentiles of the BPs of the Intersalt study.

```
quantile(intersalt$bp, probs = c(0.1, 0.25, 0.5, 0.75, 0.9))
```

```
##    10%    25%    50%    75%    90%
## 67.220 70.575 73.150 76.450 79.130
```

To specify the linear interpolation strategy recommended by the CLSI, the `type` parameter can be set to 6.

```
quantile(intersalt$bp, probs = c(0.1, 0.25, 0.5, 0.75, 0.9), type = 6)
```

```
##    10%    25%    50%    75%    90%
## 66.780 70.325 73.150 76.950 79.200
```

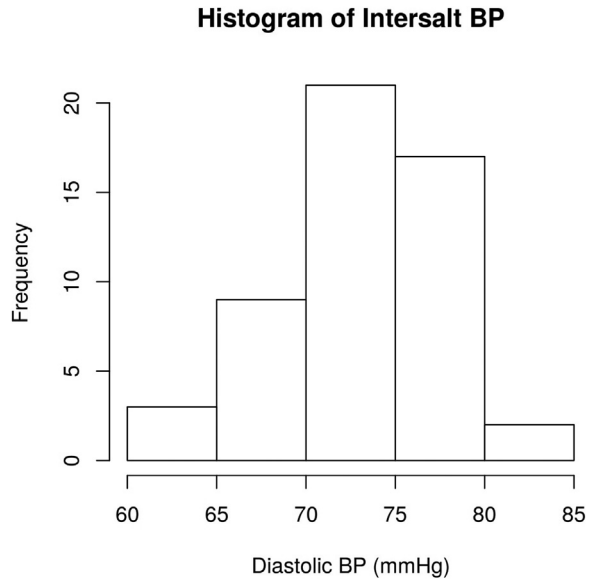


FIGURE 2.1 A histogram of the Intersalt diastolic blood pressure data.

The CLSI strategy can also be accomplished by hand by: (1) sorting the results; (2) calculating the rank of the desired quantile estimate by multiplying it by $(n + 1)$, where n is the number of observations; and (3) interpolating the desired quantile estimate from the value found from step (2).

For example, if the 95th percentile is desired and there are 52 observations, then the rank of the 95th percentile is $0.95 \times 53 = 50.35$. This means that the sample with the 50.35th rank represents the 95th percentile estimate. That is, we must interpolate a value that is 0.35 (i.e., 35%) of the way between the 50th and 51st results (taken from the list shown above) as follows:

$$79.7 + 0.35 \times (81.4 - 79.7) = 80.295.$$

If one wants to find the reference interval of a population of putatively normal values, this can be accomplished with the `quantile()` function. There is no reason (beyond compliance) to believe that the CLSI-recommended method (type = 6) is the best in all circumstances, but, in most cases, all the R methods will perform similarly.

Is my data normally distributed?

Many of the decisions we make in clinical laboratory medicine are predicated on whether the data are significantly different from the normal distribution. Most students assume that, if a histogram looks like a "bell curve," this is sufficient evidence of normality. It is not. However, if the distribution is highly skewed, then it is certainly evidence that the data are *not* normally distributed.

There are a number of simple tests to answer the question, "Is my data's distribution significantly different from the normal distribution?"

Make a histogram

The first approach is to prepare a histogram. In the R language, this is achieved with the `hist()` command (see Fig. 2.1):

```
hist(intersalt$bp,
     xlab = "Diastolic BP (mmHg)", # provide x-axis label
     main = "Histogram of Intersalt BP") # provide y-axis label
```

Prepare a normal QQ plot

The normal quantile−quantile (QQ) plot is a rapid and effective manner of gauging the normality of a distribution. The plot compares the empirical quantiles of a sampled distribution with the corresponding theoretical quantiles of the normal distribution. The closer the normal QQ plot is to the line of identity, the better the assumption of normality becomes. It is very easy to prepare in the R language, as there is function `qqnorm()` dedicated to generating it (see Fig. 2.2)

```
qqnorm(intersalt$bp)
qqline(intersalt$bp) # adds the line of identity
```

In this case, the left tail shows larger deviation from normality than the right, but the relative straightness of the normal QQ plot gives confidence that a normal assumption is reasonable.

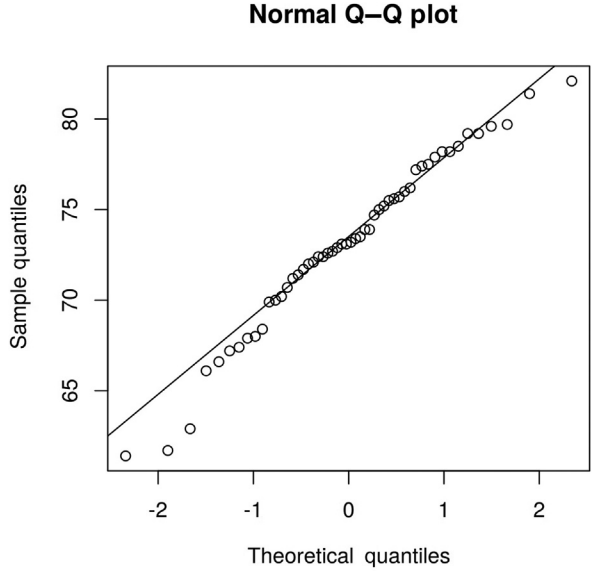


FIGURE 2.2 A normal quantile−quantile plot.

Calculate the skewness and kurtosis

The skewness is a numerical measure of a distribution's asymmetry, either leftward or rightward. A distribution is said to have a negative skew when the left tail is longer and a positive skew when the right tail is longer.

Skewness is calculated from the ratio of the "third moment" of the distribution, denoted m_3, to the cube of the SD, s, as follows: (Fig. 2.3)

$$\text{Skewness} = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{3/2}}. \quad (2.4)$$

The skewness of a normal distribution is 0. Interpretation of skewness is somewhat arbitrary, but rules of thumb are applied as follows:

- If the absolute value of the skewness is <0.5, it is considered mildly skewed.
- If the absolute value of the skewness is ≥ 0.5 and is <1, it is considered moderately skewed.
- If the absolute value of the skewness is ≥ 1, it is considered highly skewed.

We do not need to go to the trouble of using Eq. (2.4) to calculate the skewness unless we feel so determined. The calculation (including confidence intervals) is built into the R language's DescTools package and can be called as follows:

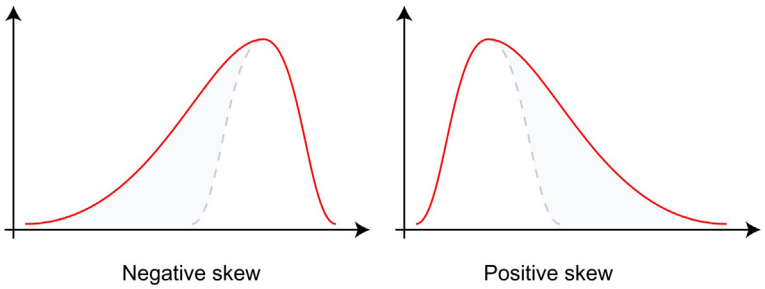


FIGURE 2.3 Negative (*left*) skew and positive (*right*) skew.

```
library(DescTools)
Skew(intersalt$bp, conf.level = 0.95)
```

```
##         skew       lwr.ci       upr.ci
## -0.46010230 -0.92059805 -0.01591163
```

Our result of −0.46 indicates only mild negative (leftward) skewing.

Kurtosis is a numerical measure of the tailedness of a distribution. It is calculated in an analogous fashion to that of Eq. (2.4) but using the ratio of the distribution's *fourth* moment to the *fourth* power of the SD. Strictly speaking, the normal distribution has a kurtosis of 3, but, for simplicity, the kurtosis is often reported as the "excess kurtosis," which is the raw value minus 3. This causes the excess kurtosis of the normal distribution to be 0, the excess kurtosis of "leptokurtic" (tall, slender, and pointy) distributions to be greater than 0, and the excess kurtosis of "platykurtic" (short, flat, and broad) distributions to be less than 0. We can calculate the excess kurtosis in R using the following:

```
Kurt(intersalt$bp, conf.level = 0.95)
```

```
##       kurt     lwr.ci     upr.ci
## -0.1618429 -0.8408011  1.1293546
```

Similar to the skewness, the excess kurtosis of −0.16 has a 95% confidence interval that includes 0, suggesting that it is not significantly different from that of a normal distribution.

Statistical tests for normality

A number of statistical tests for normality are implemented in R. Examples include the Shapiro–Wilk Test [6], its simplified variant, the Shapiro–Franci test [7], the Lilliefors (Kolmogorov–Smirnov) test [8], and the Anderson–Darling test [9]. These are presented here with references to the procedures but without the underlying calculations.

The Shapiro–Wilk test

This is implemented in R with the `shapiro.test()` function:

```
shapiro.test(intersalt$bp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  intersalt$bp
## W = 0.97341, p-value = 0.2932
```

The *P*-value indicates that the distribution is not significantly different from the normal distribution.

The Shapiro–Francia test

This function is implemented in the DescTools [10] package and is invoked as follows:

```
ShapiroFranciaTest(intersalt$bp)
```

```
##
##  Shapiro-Francia normality test
##
## data:  intersalt$bp
## W = 0.97639, p-value = 0.326
```

The *P*-value indicates that the distribution is not significantly different from the normal distribution.

The Lilliefors test

This function is implemented in the DescTools [10] package and is invoked as follows:

```
LillieTest(intersalt$bp)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  intersalt$bp
## D = 0.077987, p-value = 0.598
```

The *P*-value indicates that the distribution is not significantly different from the normal distribution.

The Anderson–Darling test

This function is implemented in the DescTools [10] package and is invoked slightly differently, as the distribution (in this case, the normal distribution) to which your data are to be compared and the associated defining parameters (in this case, the mean and the SD) need to be specified.

```
m <- mean(intersalt$bp)
s <- sd(intersalt$bp)
AndersonDarlingTest(intersalt$bp, "pnorm", mean = m, sd = s)
```

```
##
##  Anderson-Darling test of goodness-of-fit
##  Null hypothesis: Normal distribution
##  with parameters mean = 7.315e+01, sd = 4.787e+00
##
## data:  intersalt$bp
## An = 0.35745, p-value = 0.8891
```

The *P*-value indicates that the distribution is not significantly different from the normal distribution.

Detecting outliers

An outlier is a data point that differs significantly from other observations and should be considered for remeasurement or removal from the data set, since it may reflect an experimental error or contamination of the cohort with subjects who do not conform to the a priori assumptions. For example, suppose that you are performing a reference interval study for fasting serum insulin on a healthy adult population and you find one insulin result that is disproportionately high. After investigating, you find that this subject is obese and was not excluded as per study protocol. You conclude this subject's data should be removed, because they are likely insulin-insensitive. Alternatively, you might hypothesize that this subject's specimen had a heterophile interference and their sample requires remeasurement. However, unless you discover a valid reason to exclude a result based on analytical error or clinical grounds, it is *not good scientific practice* to remove outliers. Rather, readers are encouraged to use statistical tests resistant to the effects of outliers.

While a thorough discussion of identification and management of outliers is beyond the scope of this chapter, it is useful to present some of the traditionally employed strategies for finding them.

The Grubbs test

The Grubbs test [11] is based on the assumption that the data are normally distributed. The test is used to identify a single outlier, and the test statistic, G, is defined as:

$$G = \frac{\max|x_i - \bar{x}|}{s}$$

where $\bar{x}$ is the sample mean, s is the SD, and max refers to the largest value. So, for example, if we take a small data set of 10 values: $\mathbf{x} = \{-1.3, 0.7, -0.8, -0.7, 0.1, -0.3, -0.7, 0.0, 0.6, 4.2\}$, we can see that the mean is $\bar{x} = 0.18$ and $s = 1.55$, so that $G = 2.60$, which is the value associated with $x_{10} = 4.2$.

The critical value for the test statistic, G, can be found in a table for different values of α and N or calculated by using the following formula (for the two-sided hypothesis):

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

where $t_{\alpha/(2N),N-2}$ denotes the upper critical value of the student's t-distribution for $N-2$ degrees of freedom and a significance level of $\alpha/2N$. In this case, the critical G is 2.29, and the calculated value of G of 2.60 means that $x = 4.2$ is an outlier, because its value of G exceeds the critical value. A one-sided formulation of the formula above is obtained by simply replacing $\alpha/2N$ by α/N.

Grubbs' test is implemented in the R outliers' package, and the same analysis can be invoked as follows:

```
library(outliers)
x <- c(-1.3, 0.7, -0.8, -0.7, 0.1, -0.3, -0.7, 0.0, 0.6, 4.2)
grubbs.test(x)
```

```
##
##  Grubbs test for one outlier
##
## data:  x
```

```
## G = 2.59630, U = 0.16778, p-value = 0.001165
## alternative hypothesis: highest value 4.2 is an outlier
```

This gives us the same results but with some additional statistical information.

Dixon test

Dixon's Q-test [12] is another test used for the identification of outliers, which again assumes that the data are normally distributed. Suppose that we have a data set that we have *sorted* (i.e., $x_1 < x_2 < x_3 \ldots < x_N$) and the potential outlier is x_N. First, we calculate the "gap" value, Q, which is the difference between x_N and its nearest neighbor x_{N-1}:

$$Q = \frac{|x_N - x_{N-1}|}{w}$$

where w is the range of the data, that is, $w = x_N - x_1$. If w exceeds a critical value (which can only be obtained from tabulated values, not a formula), then the data point is considered an outlier. Critical values Q for $\alpha = 0.1$, 0.05, and 0.01 are provided in Table 2.2.

Considering the small data set we used above, first, we sort it to obtain this data set: −1.3, −0.8, −0.7, −0.7, −0.3, 0.0, 0.1, 0.6, 0.7, 4.2. The value in question is 4.2. Its nearest neighbor is 0.7. The calculated value of Q is:

$$Q = \frac{|4.2 - 0.7|}{4.2 - (-1.3)} = 0.64.$$

This value $Q = 0.64$ exceeds the critical value of 0.466 for 10 observations and $\alpha = 0.05$, as shown in Table 2.2.

A slightly more sophisticated variant Dixon's Q-test taking into account the sample size [13] is implemented in the R outliers' package, and the same analysis can be invoked as follows:

```
library(outliers)
x <- c(-1.3, 0.7, -0.8, -0.7, 0.1, -0.3, -0.7, 0.0, 0.6, 4.2)
dixon.test(x)
```

```
##  Dixon test for outliers
##
## data:  x
## Q = 0.7, p-value < 2.2e-16
## alternative hypothesis: highest value 4.2 is an outlier
```

The reader will notice that the calculated value of Q is different than the above, which is attributable to the slightly more sophisticated approach used.

Tukey's fences

Another way to flag a potential outlier was suggested by Tukey [14]. If a result is more than 1.5 IQRs below the first quartile, Q_1, or more than 1.5 IQRs about the third quartile, Q_3, then the result can be considered an outlier. For the data set we have been considering in this section, the first quartile can be calculated to be $Q_1 = -0.700$, the third quartile to be $Q_3 = 0.475$, and the IQR to be IQR = 1.175. The Tukey fences are therefore [−2.4625, 2.2375]. Because 4.2 lies outside this interval, it is considered an outlier.

Chauvenet's criterion

Chauvenet's criterion [15] is another means of identifying an outlier in a normally distributed data set. The principle behind the test is that, if a value falls outside a probability band centered at the mean of the data, then it can be considered an outlier. Despite its frequent application in the field of clinical chemistry, its use is not recommended because of errors in its derivation; it was considered antiquated some 60 years ago [16]. The reader determined to use Chauvenet's criterion may do so as follows.

A measurement can be rejected if its probability of occurrence is less than $1/2N$ [15]. For the data we have been considering in this section, $N = 10$, the mean is 0.180, the SD is 1.548, and the z-score of the prospective outlier, $x = 4.2$, is 2.596. The minimum acceptable probability is $1/2N = 1/20 = 0.05$. Any observation having a probability of occurrence less than this is considered an outlier.

From the perspective of how z-tables for the standard normal distribution are tabulated (i.e., presenting the area under the distribution function from $-\infty$ to z), this corresponds to a z-score for a probability of $1/4N$. In this case, for the data set we have been evaluating, the critical $z = 1.960$. Because our calculated value for $z = 2.596$ exceeds 1.960, we can say $x = 4.2$ is an outlier.

In the R language, the Chauvenet's criterion can be calculated with the `qnorm()` function for any value of N

TABLE 2.2 Critical values for Dixon's Q statistic.

Number of values	3	4	5	6	7	8	9	10
Q90%:	0.941	0.765	0.642	0.560	0.507	0.468	0.437	0.412
Q95%:	0.970	0.829	0.710	0.625	0.568	0.526	0.493	0.466
Q99%:	0.994	0.926	0.821	0.740	0.680	0.634	0.598	0.568

by calculating `qnorm(1 - 1/(4*N))`. For example, if $N = 30$, the critical z is calculated as follows:

```
N <- 30
qnorm(1-1/(4*N))

## [1] 2.39398
```

Those invoking Chauvenet's method will often apply it serially, removing outliers until no samples fail the criterion. This is not sound practice.

Common inferential statistics

This section will not cover the theoretical basis for these tests. This is covered in all introductory statistical textbooks, some of which are available online [17,18]. Rather, this section will inform the reader when each type of test is appropriate, how to perform the test, how to interpret the results, and how to perform the test in the R language.

The two-sample *t*-test

The two-sample *t*-test is a way of assessing whether the means of two populations are different. In statistical terms, it is more correct to say, "the test allows us to accept or reject the null hypothesis," where the null hypothesis is the hypothesis that the means are equal.

The *t*-test, in its simplest form, compares the means of two unrelated populations. For example, if you were comparing the mean mid-winter plasma vitamin D concentration in a cohort of Icelanders to Floridians (who are in no way matched), you would use a two-sample *t*-test.

The assumption of the *t*-test is that the quantity being compared in the two populations is normally distributed. One may wish to assume that the variances of the two populations are equal or not, but the ensuing calculations will be different accordingly. If the reader wishes to assess the normality of the distributions, strategies in the section "Is my data normally distributed?" can be applied. If the data are found not to conform to a normal distribution, normalizing transformations such as the logarithm, Box−Cox [19], or one of many others available in R [20] can be employed first. Finally, the data should be sampled independently from the two populations.

Examples

Suppose that we have a small sampling of mid-winter plasma vitamin D levels from Icelandic and Floridian adult males (see Table 2.3) that we wish to compare with a *t*-test.

In the R language, a *t*-test is accomplished as follows:

```
t.test(VitD$Iceland, VitD$Florida)

##
##  Welch Two Sample t-test
##
## data:  VitD$Iceland and VitD$Florida
## t = -12.432, df = 21.113, p-value = 3.516e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -40.00254 -28.54069
## sample estimates:
## mean of x mean of y
##  51.77245  86.04407
```

The output indicates that the difference between the two means, 51.8 and 86.0, is statistically significant based on the *P*-value since $P \ll .05$.

If a one-sided *t*-test is preferred because there is an a priori assumption that the Icelandic vitamin D levels should be lower, then the code is modified as follows:

```
t.test(VitD$Iceland, VitD$Florida, alternative = "less")
```

TABLE 2.3 Independently sampled plasma vitamin D levels from Icelandic and Floridian adult males.

Iceland VitD (nmol/L)	Florida VitD (nmol/L)
52.6	85.0
68.3	84.9
46.6	85.2
59.0	76.8
54.9	81.4
37.4	94.6
50.2	94.9
60.9	82.2
48.7	81.9
39.2	92.1
58.6	88.0
46.4	90.1
51.7	80.2
37.6	87.3
64.6	91.3
–	81.8
–	85.0

The paired *t*-test

The results are similarly significant to the *t*-test.

Wilcoxon signed-rank test

This is the nonparametric equivalent of the paired *t*-test. It makes no assumption about the distributions of the original population themselves, but it does assume that the distributions of the *differences* are at least symmetrical (though there is no need for the differences to be normally distributed). In the R language, the syntax to obtain the Wilcoxon signed-rank test on paired data sets, x_1 and x_2, is:

If the data are paired, that is, one is comparing populations before and after some intervention or the populations respectively comprise cases and matched controls, then the appropriate test is a paired two-sample *t*-test. The test assumes that the distributions of the *differences* between the paired values of the two populations are normally distributed rather than the two distributions themselves. So, if we were comparing the vitamin D levels of a single group of Icelanders in the summer versus winter, we would use a paired *t*-test. The syntax for calling a paired *t*-test on two populations, x_1 and x_2, in R is similarly straightforward:

```
t.test(x1, x2, paired = TRUE)
```

Wilcoxon rank-sum test (Mann–Whitney *U*-test)

This is the nonparametric equivalent of the *t*-test and as such does not require that the populations compared be normally distributed. Applying this test to the vitamin D data, as shown in Table 2.3, and using the R language implementation of this test, we have:

```
wilcox.test(VitD$Iceland, VitD$Florida)
```

```
##
##  Wilcoxon rank sum test
##
## data:  VitD$Iceland and VitD$Florida
## W = 0, p-value = 3.535e-09
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(x1, x2, paired = TRUE)
```

The chi-square test

The chi-square test is used to test for statistically significant differences in frequencies observed in different categories.

Application to proportions

For example, if we had four clinical chemistry training programs and we knew how many students had passed and failed over the course of the last 20 years, we could look for

TABLE 2.4 Relationship between smoking frequency (heavy, regular, occasional, and never) and exercise frequency (frequent, some, and none) in survey data from the MASS package in R.

	Heavy	Never	Occas	Regul
Freq	7	87	12	9
None	1	18	3	1
Some	3	84	4	7

significant differences in the pass/fail frequencies to make evidence-based recommendations to our younger colleagues.

	School A	School B	School C	School D
Pass	35	33	34	25
Fail	4	3	5	11

In the R language we can construct a data frame or matrix to house the data and apply the chi-square test as follows:

```
chisq.test(cbind(pass,fail))
```

```
##
##  Pearson's Chi-squared test
##
## data:  cbind(pass, fail)
## X-squared = 8.7483, df = 3, p-value = 0.03283
```

The resulting *P*-value of 0.03 indicates that school D, aptly named, should not be on the recommendation list.

Another R function, `prop.test()`, is specifically tailored for testing proportions. Not surprisingly, it gives the same result:

```
prop.test(cbind(pass,fail))
```

```
##
##  4-sample test for equality of proportions without continuity
##  correction
##
## data:  cbind(pass, fail)
## X-squared = 8.7483, df = 3, p-value = 0.03283
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4
## 0.8974359 0.9166667 0.8717949 0.6944444
```

General application to contingency tables

If you have a more general contingency table, you can apply the chi-square test similarly. For example, the MASS package in the R language [21] has a data set containing survey responses from 237 statistics students from the University of Adelaide. A contingency table for smoking and exercise can be prepared from this data set: (Table 2.4)

Applying the `chisq.test()` function directly to this contingency table gives:

```
library(MASS) #load the package with the data
data(survey) #load the data
cont.table <- table(survey$Exer, survey$Smoke) #prepare table
chisq.test(cont.table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont.table
## X-squared = 5.4885, df = 6, p-value = 0.4828
```

These results indicate that there is no significant relationship between exercise frequency and heaviness of smoking.

Methods of regression

Ordinary least squares

In laboratory medicine, regression is used to prepare calibration curves, assess linearity, and compare methods. Usually, we are discussing linear regression with a single explanatory (independent) variable, but regression can be mulitvariate and it can be nonlinear. For the moment, we will confine ourselves to discussing the simplest regression method, ordinary least squares (OLS) linear regression. The method, usually attributed to Carl Friedrich Gauss, works by finding the line that minimizes the sum of squared residuals of a sequence of points (see Fig. 2.4).

The residuals, in this case, are the vertical distances between the points and the line. The minimization of these squared distances leads to a regression line whose slope and intercept can easily be calculated.

OLS regression assumes that there is *no error* in the independent (i.e., *x*-axis) variable. This is never true in bioanalytical chemistry, of course. However, there are settings where the error in the *x*-axis variable is low relative to that of the *y*-axis variable. For example, if you are doing a method comparison against an absolute reference method, it might be reasonable to assume no error in the *x*-axis. Another example is a calibration curve, where the assigned concentrations of the calibrators are considered correct but the instrument responses on the *y*-axis are subject to error. OLS regression also assumes that the scatter about the regression line is constant and not a function of *x*. This property of constant error is called "homoscedasticity," and it is likewise almost never true (see Fig. 2.5). If the data exhibit obvious heteroscedasticity, weighting is often applied to the regression (to be discussed).

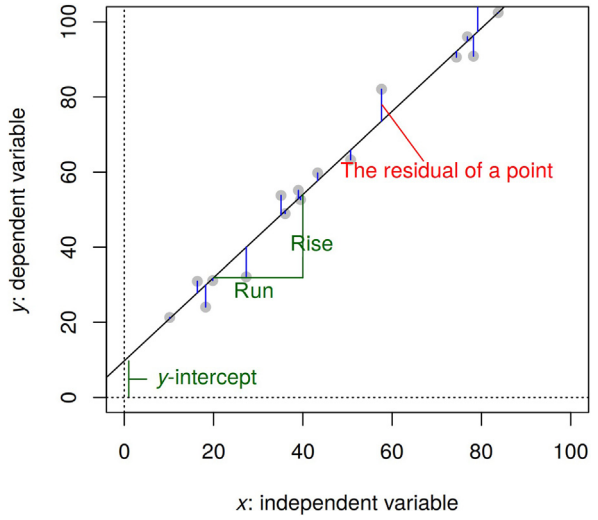


FIGURE 2.4 Typical scatterplot with the residuals of the points shown as vertical blue line segments. The ordinary least squares regression line is shown in solid black.

Computational methods for OLS regression report back a number of useful parameters and diagnostics for us to examine. These are usually the slope, the *y*-intercept, the residual standard error (RSE), and the coefficient of determination, R^2.

The slope is a measure of the regression line's steepness and is calculated from the ratio of the rise to the run (see Fig. 2.4). The *y*-intercept is the *y* value of the intersection of the regression line and the *y*-axis (i.e., the line $x = 0$). If the values on the *x*- and *y*-axes are measurements of the same quantity using two different analytical methodologies, the nonzero *y*-intercept indicates that the *y*-axis method has a constant bias compared with the *x*-axis method, and the fact that the slope is not equal to 1 indicates that the *y*-axis method has a proportional (i.e., *x*-dependent) bias compared with the *x*-axis method.

The RSE is given by:

$$\mathrm{RSE} = \sqrt{\frac{\sum_{i=1}^{n} \mathrm{resid}_i^2}{n-2}} \tag{2.5}$$

where resid_i is the residual of the *i*th value of *x*. The RSE therefore gives a measure of the spread of the residuals and is a goodness of fit measure of the linear regression line to the data.

The coefficient of determination is the square of the Pearson correlation coefficient, *r*, where *r* is defined as:

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \\ &= \frac{\sigma_{xy}}{\sqrt{\sigma_x^2\sigma_y^2}} \\ &= \frac{\sigma_{xy}}{\sigma_x\sigma_y} \end{aligned} \tag{2.6}$$

The numerator of *r*, denoted σ_{xy}, is referred to as the *covariance* of *x* and *y*. The correlation coefficient is a

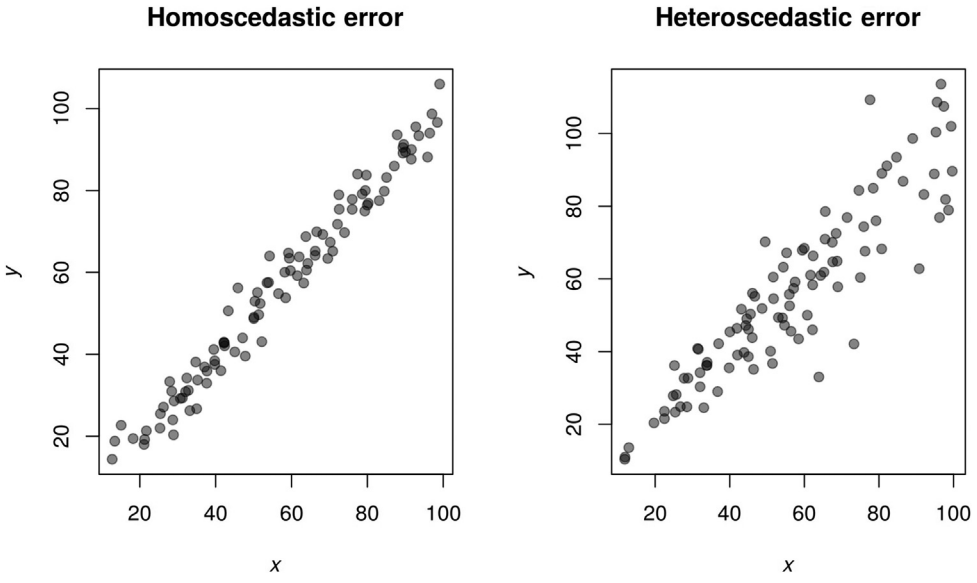


FIGURE 2.5 Examples of data showing homoscedastic error (*left*) and heteroscedastic error (*right*).

measure of how close two variables are to having a linear relationship with one another. It can take on values between -1 and $+1$. If the value is positive, the OLS regression line will have a positive slope, and, if the value is negative, the regression line will have a negative slope. This is because the OLS slope can be expressed as $\beta = r \times \sigma_y/\sigma_x$, meaning that r and β have the same sign since both σ_x and σ_y are positive. A value of $r = 0$ implies that the data are uncorrelated.

The correlation coefficient is an oft-abused quantity in methodological comparisons, because it can be inflated by the inclusion of an outlier that lies close the regression line, thus imparting the illusion of "good correlation." The author has even seen manuscripts where said outlier is excluded from the scatterplot and the inflated value of r is ascribed to the remaining visible data. Fig. 2.6 shows an extreme example of this phenomenon.

In the R language, the OLS regression is built in as the `lm()` function. Using the data shown in Fig. 2.4 and Table 2.5, the summary of the results of the linear model is obtained as follows:

```
x <- c(89.0,79.2,35.1,57.6,96.7,98.2,18.2,16.4,39.5,43.3,
       74.4,78.2,10.2,76.8,27.3,50.7,39,19.8,36,83.8)
y <- c(110.7,107.7,53.8,82.1,112.8,112.8,24,30.9,52.7,
       59.8,90.6,90.9,21.3,96,32.1,63.3,55.1,31.1,48.9,102.5)
linmod <- lm(y ~ x)
# the notation y ~ x means y is a function of x
summary(linmod)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
## Min 1Q Median 3Q Max
## -7.8589 -2.9527 -0.3047 2.2277 10.2365
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.71087 2.31172 4.201 0.000537 ***
## x 1.10799 0.03827 28.951 < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 4.809 on 18 degrees of freedom
## Multiple R-squared: 0.979, Adjusted R-squared: 0.9778
## F-statistic: 838.2 on 1 and 18 DF, p-value: < 2.2e-16
```

TABLE 2.5 Example correlation data.

X	*Y*
89.0	110.7
79.2	107.7
35.1	53.8
57.6	82.1
96.7	112.8
98.2	112.8
18.2	24.0
16.4	30.9
39.5	52.7
43.3	59.8
74.4	90.6
78.2	90.9
10.2	21.3
76.9	96.0
27.3	32.1
50.7	63.3
39.0	55.1
19.8	31.1
36.0	48.9
83.8	102.5

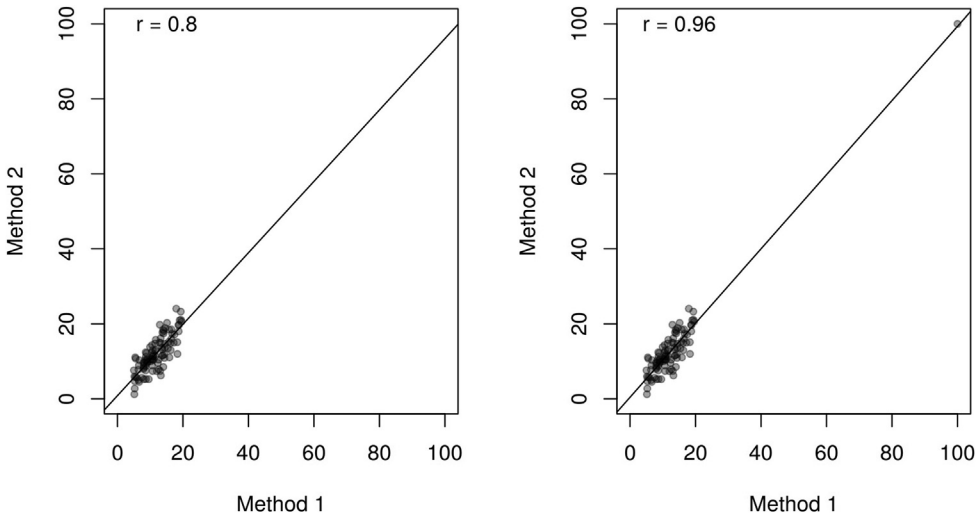


FIGURE 2.6 The spuriously positive effect of an outlier on the correlation coefficient.

Among a number of other results, we can see that the intercept of the regression line is 9.71, and the slope of the regression line is 1.11. The RSE is 4.81, and $R^2 = 0.979$.

Weighted ordinary least squares

As mentioned, the assumption of OLS regression is that the data show a homoscedastic error. When this is not the case, there is often a desire to limit the effect of the "spread" of x results. This can be accomplished by applying weights to the residuals in the calculation process, which can be chosen to make the regression line to be more affected by the lower values of x and less affected by the higher values of x. This is also common practice in calibration curve preparation, where accuracy of results in the concentration domain of the low calibrators is more important than that of higher calibrators.

Typically, one can weight with $1/x$ or $1/x^2$ weights. This is accomplished in the R language very easily. A small set of mock data is created:

```
xh <- c(1,3,4,6,7,8,10)
yh <- c(1.2,2.9,5.5,7.2,6.2,10,8.1)
summary(lm(yh ~ xh, weights = 1/xh))
```

```
##
## Call:
## lm(formula = yh ~ xh, weights = 1/xh)
##
## Weighted Residuals:
##       1       2       3       4       5       6       7
## -0.1496 -0.2470  0.5915  0.3693 -0.4099  0.6103 -0.6805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3604     0.6044   0.596 0.576941
## xh            0.9892     0.1408   7.023 0.000903 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5609 on 5 degrees of freedom
## Multiple R-squared:  0.908,  Adjusted R-squared:  0.8896
## F-statistic: 49.33 on 1 and 5 DF,  p-value: 0.0009028
```

The effect of the weighting is best appreciated graphically (see Fig. 2.7).

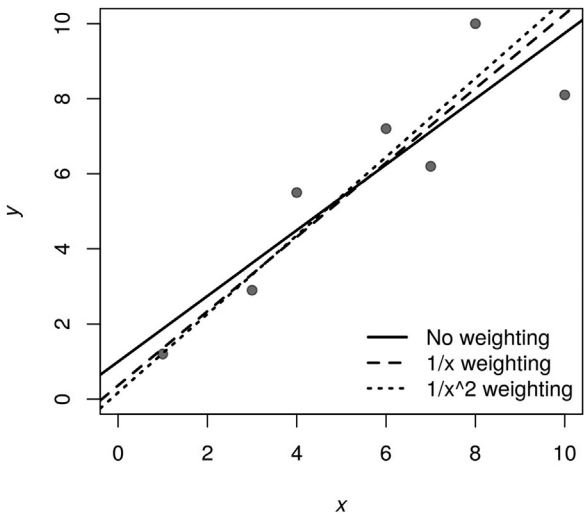


FIGURE 2.7 The effect of weighting an ordinary least squares regression of heteroscedastic data.

Deming regression

Some of the deficiencies of OLS regression are addressed by the method of Deming [22,23]. Specifically, Deming regression does not make the assumption that the x-axis variable has no error. Rather, the data of both the x- and y-axes are considered to have error associated with them. Accordingly, Deming regression calculates its residuals as the *perpendicular* distance to the regression line rather

than the vertical distance, as shown in Fig. 2.8. The Deming approach assumes that the error (i.e., scatter) is independent between the variables (which should be method comparison studies) and that it is normally distributed. However, Deming regression requires prior knowledge of the ratio of the variances of the error. That is, it does not assume that the variables x and y have the same analytical coefficient of variation. Unless specified, the ratio of the variance is assumed to be 1, and this may not be a correct assumption.

A number of packages that can perform Deming regression are available for the R language. These include the mcr [24] package from Roche Diagnostics and the deming package from Mayo Clinic [25]. Using the mcr package, the `mcreg()` function generates the Deming regression results as follows, assuming a value of 1 for the ratio of variances, as expressed in the parameter setting `error.ratio = 1`:

```
library(mcr)
x <- c(89.0,79.2,35.1,57.6,96.7,98.2,18.2,16.4,39.5,43.3,
       74.4,78.2,10.2,76.8,27.3,50.7,39,19.8,36,83.8)
y <- c(110.7,107.7,53.8,82.1,112.8,112.8,24,30.9,52.7,59.8,
       90.6,90.9,21.3,96,32.1,63.3,55.1,31.1,48.9,102.5)
deming <- mcreg(x,y,method.reg = "Deming", error.ratio = 1)
printSummary(deming)
```

```
##
##
## ----------------------------------------
##
## Reference method: Method1
## Test method: Method2
## Number of data points: 20
##
## ----------------------------------------
##
## The confidence intervals are calculated with
## bootstrap ( quantile ) method.
## Confidence level: 95%
## Error ratio: 1
##
## ----------------------------------------
##
## DEMING REGRESSION FIT:
##
## EST SE LCI UCI
## Intercept 9.006000 NA 4.985596 13.150060
## Slope 1.121171 NA 1.052231 1.205403
##
##
## ----------------------------------------
##
## BOOTSTRAP SUMMARY
##
## global.est bootstrap.mean bias bootstrap.se
## Intercept 9.00600 9.03326 0.02726 2.13112
## Slope 1.12117 1.12205 0.00088 0.04032
##
## Bootstrap results generated with environment RNG
```

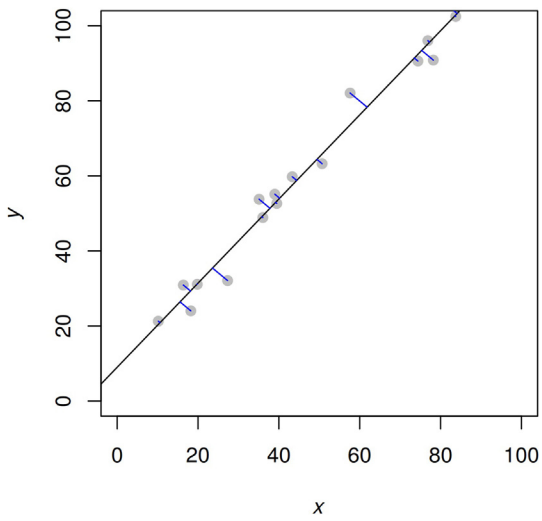


FIGURE 2.8 Perpendicular residuals of Deming regression.

```
wdeming <- mcreg(x,y,method.reg = "WDeming", error.ratio = 1)
```

Deming regression can also be weighted for the same motivations that weighting is performed in OLS regression. Weighting is specified by choosing the regression setting the parameter `method.reg = "WDeming"`:

```
x <- c(1,3,4,6,7,8,10,8)
y <- c(1.2,2.9,5.5,7.2,6.2,10,8.1,1)
pb <-  mcreg(x, y, method.reg = "PaBa")
printSummary(pb)
```

In this case, a slightly different intercept and slope of 8.01 and 1.14 are produced.

Passing–Bablok regression

The reader is undoubtedly aware that both OLS and Deming regression are unduly affected by outlier points. There exists a class of regression methods resistant to the effect of a small number of outliers. These are called "robust" regression methods, and Passing–Bablok (PB) regression is among them [26–28]. The general approach of PB regression is to identify all possible pairs of points in the data set, calculate a slope for each pair, and use the median as the slope of the regression line. The intercept is then calculated in a mathematically consistent manner. No assumptions are made about errors in the x or y variables, and there are no assumptions about the distribution of the error (i.e., scatter) about the regression line. The problem with PB regression is that it is

computationally intensive in that when there are N data points, there are $\binom{N}{2} = \frac{N(N-1)}{2!}$ possible pairs of slopes, making the calculation quite slow for large data sets. While there are many strategies for robust regression, it is a peculiarity of Clinical Chemistry as a field that it is culturally devoted to this particular method. Like Deming regression, PB regression is available in the R language in both the mcr and deming packages. To illustrate its resistance to the effect of an outlier, we can introduce the outlier point $(x, y) = (8, 1)$ into the data set used in Fig. 2.7 and compare the effect on PB regression, OLS regression, and Deming regression. These are calculated in R in a manner analogous to Deming regression:

```
library(pROC)
data(aSAH)
ndka.roc <- roc(outcome ~ ndka, data = aSAH)
plot(ndka.roc)
```

```
ndka.roc
```

```
##
## Call:
## roc.formula(formula = outcome ~ ndka, data = aSAH)
```

The regression coefficients of the three approaches are shown in Table 2.6 and illustrated in Fig. 2.9.

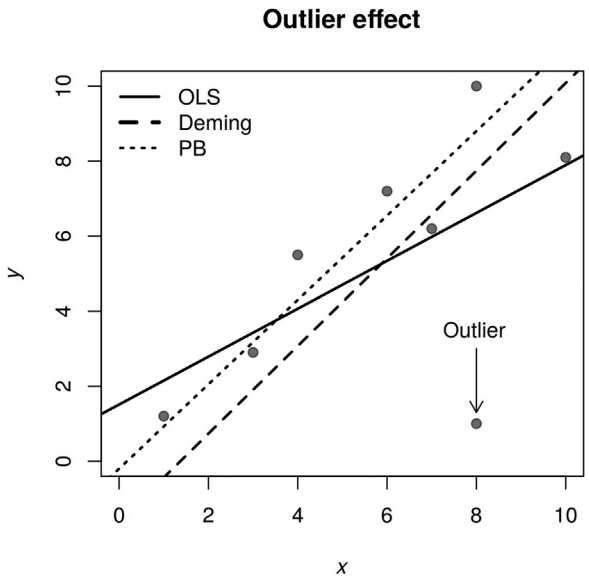


FIGURE 2.9 Illustration of the effect of an outlier on ordinary least squares, Deming, and PB regression.

Evaluation of diagnostic test performance

It is common to discuss sensitivity of an assay from an analytical perspective. That is, "How low can the assay measure?" Likewise, we talk about assays from the perspective of their analytical specificity. That is, "What are the known cross-reacting substances and how much do they interfere (%) on a molar basis?"

We can also discuss the sensitivity and specificity of tests from an epidemiological perspective, which addresses assay performance from a *diagnostic* perspective rather than an *analytical* perspective. The calculation of sensitivity and specificity requires knowledge of the true disease state of the individual by some other means (gold-standard testing or clinical criteria).

Sensitivity

The sensitivity of a test can be remembered by the brief phrase "positivity in disease." More accurately, *the sensitivity of a diagnostic test is the probability of obtaining a positive result in a diseased individual.*

- The sensitivity of a diagnostic test is an intrinsic property of the diagnostic test and the diagnostic threshold defined by the user.
- A test that has perfect (100%) sensitivity catches *all* diseased individuals but does not necessarily catch *only* diseased individuals.

Mathematically, the sensitivity is defined as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.7}$$

where TP is the number of true positives and FN is the number of false negatives. The reader may convince themselves that this is indeed the probability of a positive result when a test is applied to a diseased individual.

Graphically, if we imagine that healthy individuals have a different distribution of possible results and diseased individuals have a distribution of possible results, we may visualize Eq. (2.7) as shown in Fig. 2.10.

Specificity

The specificity of a test can be remembered by the brief phrase "negativity in health." More accurately, *the specificity of a diagnostic test is the probability of obtaining a negative result in a healthy individual.*

- The specificity of a diagnostic test is an intrinsic property of that test and the diagnostic threshold that is defined by the user.
- A test that has perfect (100%) specificity catches *only* diseased individuals but does not necessarily catch *all* diseased individuals.

Mathematically, the specificity is defined as:

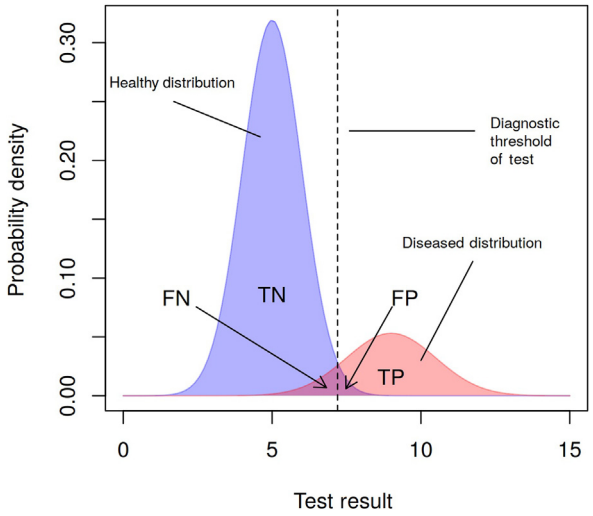


FIGURE 2.10 True/false positives and true/false negatives visualized as areas of overlapping normal distributions of test results in the diseased and healthy populations. The diagram assumes that test results higher than the diagnostic threshold are positive and that test results lower than the diagnostic threshold are negative. The true negatives are represented by the area of the healthy distribution (*blue*) that lies to the *left* of the diagnostic threshold level. The true positives are represented by the area (*pink*) to the *right* of the diagnostic threshold level. The false negatives are represented by the small portion of the diseased distribution to the left of the diagnostic threshold, and the false positives are represented by the small portion of the healthy distribution that lies to the right of the diagnostic threshold.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{2.8}$$

Positive predictive value

The positive predictive value (PPV) of a diagnostic test is *the probability that a subject who tests positive is diseased*. This is represented mathematically as:

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2.9}$$

The PPV is *not* an intrinsic feature of the diagnostic test. Rather, it is dependent on the population to which the test is applied. When the prevalence of the disease in question is low, the PPV will be correspondingly low. As the prevalence approaches 0, the PPV will also approach 0. Using the fact that $\text{Prevelance} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$, Eq. (2.9) can be reexpressed as follows:

$$\text{PPV} = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})} \tag{2.10}$$

which demonstrates that, as the prevalence goes to 0, the PPV also goes to 0. The consequence of Eq. (2.10) is that screening tests provide little value when applied to populations of low disease prevalence, because most positives will be false positives.

Negative predictive value

The negative predictive value (NPV) of a diagnostic test is *the probability that a subject who tests negative is healthy*. This is represented mathematically as:

$$\text{PPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}. \tag{2.11}$$

The NPV is gauge of the reassurance that can be provided to a patient who tests negative.

Likelihood ratios

There is an alternative manner of expressing the concepts from the section "Evaluation of diagnostic test performance," which is that of likelihood ratios. The first required concept is one of *odds*. Most readers will have a sense from colloquial usage that probability is expressed as a number between 0 and 1, while odds are expressed by numbers that may be greater than 1. The relationship between odds and probability is as follows. If the probability of an event's occurrence is p_e, then the odds for that event's occurrence, o_e, is:

$$o_e = \frac{p_e}{1 - p_e}. \tag{2.12}$$

It can be shown that, if one defines the positive likelihood ratio for a diagnostic test, denoted LR^+, as follows:

$$\text{LR}^+ = \frac{\text{sensitivity}}{1 - \text{specificity}} \tag{2.13}$$

then there is a relationship between the pretest odds of disease and the posttest odds of disease given by:

$$o_{\text{post}} = \text{LR}^+ \times o_{\text{pre}}. \tag{2.14}$$

Eq. (2.14) is a simplified formulation of *Bayes' theorem* and has very important implications for the use of diagnostic tests. The pretest odds, o_{pre}, is directly calculated by using Eq. (2.12) using the pretest probability of disease (i.e., the prevalence) *in the population to be tested*. The posttest odds of having disease is then simply a multiple of the pretest odds, which underscores the notion that if the pretest odds (and therefore the pretest probability) are low, then the odds (and probability) of finding a diseased individual using the test are likewise low.

TABLE 2.6 Comparison of outlier effects on commonly employed regression methods.

Method	Slope	Intercept
Least squares	0.6	1.5
Deming	1.2	1.6
Passing–Bablok	1.1	0.2

As a rule of thumb, if $LR^+ > 10$, the test has a large effect on increasing the probability identifying a diseased individual, if $5 \leq LR^+ \leq 10$, the test has moderate effect on increasing this probability, and if $LR^+ < 5$, the test is weak effect on increasing this probability.

Similarly, we can formulate an analogous version of Eq. (2.14) for negative tests:

$$LR^- = \frac{1 - \text{sensitivity}}{\text{specificity}} \tag{2.15}$$

and

$$o_{\text{post}} = LR^- \times o_{\text{pre}} \tag{2.16}$$

where Eq. (2.16) determines the posttest odds of being *healthy* based on the pretest odds of health.

Example calculation

Let us suppose that we have developed a diagnostic test, which we can call *my patented biomarker* (MPB), to determine whether a patient has irritable bowel syndrome (IBS) versus inflammatory bowel disease (IBD). A positive test indicates the presence of IBD. Suppose that MPB testing is performed on a cohort of patients who have had colonoscopies and long-term clinical evaluation, so that they are correctly clinically classified as either IBS or IBD. Suppose that the MPB results on a cohort of 117 people with chronic diarrhea were as follows:

From Table 2.7, we can infer that TP = 17, FN = 5, FP = 3, and TN = 92. This means that the prevalence of the disease in the testing population is Prev = (TP + FN)/(TP + TN + FP + FN) = 0.188. Therefore, using Eq. (2.12), we can calculate the positive pretest odds of IBD as $o^+_{\text{pre}} = \frac{\text{prev}}{1 - \text{prev}} = 0.232$. From Eq. (2.7), the sensitivity is found to be Sens = TP/(TP + FN) = 0.773, and from Eq. (2.8), the specificity is found to be Spec = TN/(TN + FP) = 0.968. This means that the positive likelihood ratio can be calculated:

$$LR^+ = \frac{\text{sens}}{1 - \text{spec}} \simeq \frac{0.773}{1 - 0.968} = 24.2$$

TABLE 2.7 A table of test performance of a blood biomarker my patented biomarker for inflammatory bowel disease in a mixed cohort of inflammatory bowel disease and irritable bowel syndrome.

	IBD+	IBD
MPB+	17	3
MPB	5	92

and from this, we can determine the positive posttest odds:

$$o^+_{\text{post}} = LR^+ \times o^+_{\text{pre}} = 24.2 \times 0.232 = 5.61.$$

Because the PPV corresponds to the disease prevalence in the population of patients who test positive, by substituting $o^+_{post} = 5.61$ into Eq. (2.12) and solving for the corresponding probability, the reader should be able to convince themselves that the PPV = 5.61/6.61 = 0.849, which matches the value calculated directly by using Eq. (2.9).

Receiver operating characteristic curves

The tacit assumption of the section "Likelihood ratios" was that the appropriate cutoff of a diagnostic test is already established. However, when we are evaluating a new diagnostic test which generates a numerical (and not dichotomous) result, we do not know the cutoff that will produce the best diagnostic accuracy. This is where receiver operating characteristic (ROC) curves can be used. These serve to both determine the appropriate concentration above which (or below which) to declare tests as "positive" and also give objective measure as to the test's diagnostic utility via the area under the ROC curve.

The ROC curve is a plot of sensitivity (y-axis) versus 1 − specificity (x-axis) for all possible diagnostic cutoffs. Considering Fig. 2.10, if the vertical dashed line representing the diagnostic threshold of the test was moved to a value of 5, the TN rate would decrease by about 50%, while the FN result would go to nearly 0. The FP rate would go up sharply, and the TP rate would be nearly 1. Naturally then, the sensitivity and the specificity would change. If we slide the vertical dashed line through all possible values of the test result and plot sensitivity versus 1 − specificity for all of these values, the ROC curve is obtained.

Interpreting a receiver operating characteristic curve

When we apply a diagnostic test to a population of healthy and diseased individuals, the more separated the

two distributions are, the better the diagnostic test will perform. In a situation where the healthy and diseased are well-separated, the ROC curve will look rectangular and will have an area under the curve (AUC) approaching 1. In contrast, if the test has no discriminating power between the healthy and diseased populations, the ROC curve will approach the line $y = x$ and will define an area close to $\frac{1}{2}$. Examples of the effect of population overlap on the corresponding ROC are shown in Fig. 2.11.

For this reason, ROC curves that have larger area indicate better discriminating power of the test, and better clinical test utility is expected. Obviously with any real experimental data, there will be a margin of error in the AUC, and this should be considered when comparing the AUCs of different tests.

The optimal diagnostic threshold of the test is generally defined as the value that will optimize the sum of specificity and sensitivity. Geometrically, this will be the point on the ROC curve that is furthest from the light gray line of identity shown in the ROC curves of Fig. 2.11.

There are no rules about what AUC is "good enough" but as a rule of thumb:

- $AUC > 0.9$ is excellent.
- $0.8 \leq AUC < 0.9$ is good.
- $0.7 \leq AUC < 0.8$ is fair.
- $AUC < 0.7$ is poor.
- $AUC = 0.5$ has no discriminating power.

Preparing a receiver operating characteristic curve

ROC curves are tedious to generate in spreadsheet programs but have been implemented in a number of packages for the R language.

By way of example, the R package pROC [29] provides a mock data set for 113 patients with an aneurysmal subarachnoid hemorrhage using s100 calcium-binding

TABLE 2.8 Representative data from the aSAH data set from the pROC package showing the relationship bteween s100b, ndka, and clinical outcome after aneurysmal subarachnoid hemorrhage.

Subject	Outcome	s100b	ndka
1	Good	0.13	3.01
2	Good	0.14	8.54
3	Good	0.10	8.09
4	Good	0.04	10.42
5	Poor	0.13	17.40
6	Poor	0.10	12.75
7	Good	0.47	6.00
8	Poor	0.16	13.20
9	Good	0.18	15.54
10	Good	0.10	6.01

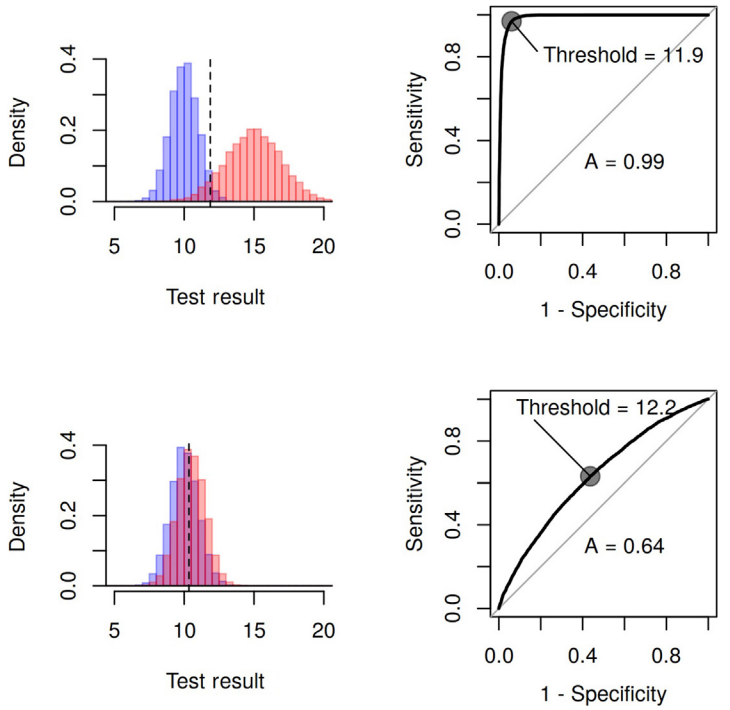


FIGURE 2.11 A comparison of receiver operating characteristic curves generated from a test with good discriminating power (*top*) between healthy (*blue*) and diseased (*red*) populations and poor discriminating power (*bottom*). The optimal test threshold that maximizes the sum of sensitivity and specificity is identified on each ROC curve and by vertical dashed lines on histograms.

protein B (s100b) and nucleoside diphosphate kinase A (ndka) as biomarkers of good or poor outcome. Relevant columns of the first 10 rows of the data set (Table 2.8) give the reader an understanding of how it must be structured to prepare an ROC curve.

The package makes preparing an ROC curve a trivial matter. Here is the code to create a curve based on ndka:

```
##
## Data: ndka in 72 controls (outcome Good) < 41 cases (outcome Poor).
## Area under the curve: 0.612
```

Fig. 2.12 shows the resulting ROC curve. The AUC, in this case 0.61, is calculated in the output among many other statistics discussed in the package documentation [29]. More elaborate ROC representations can be prepared with ease. For example, Fig. 2.13 shows the s100b threshold providing optimal diagnostic accuracy, the optimized sensitivities, and specificities with their confidence intervals, and a confidence band on the ROC.

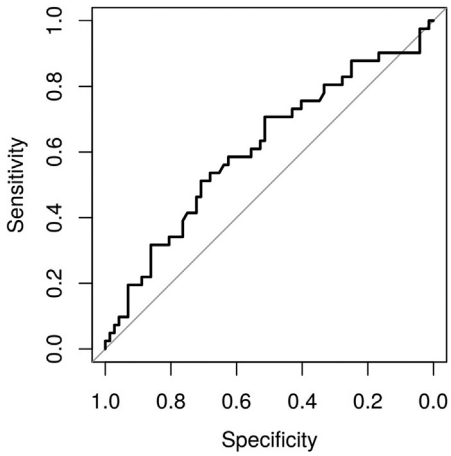


FIGURE 2.12 A basic receiver operating characteristic curve prepared by the roc( ) function and in the pROC package.

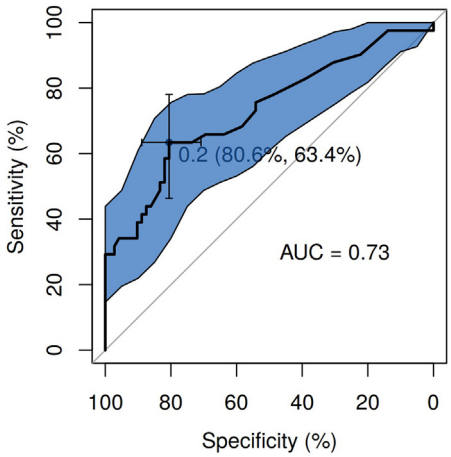


FIGURE 2.13 Receiver operating characteristic curve for biomarker S100b to predict poor or good outcome after subarachnoid hemorrhage.

Suggested additional topics for study

- Method validation statistics
- Quality control statistics
- Dealing with missing values
- Traceability and commutability
- Biological variation and reference change values
- Total allowable error and uncertainty in the measurement
- Multivariate regression
- Nonlinear regression
- The propagation of error
- Analysis of variance and Kruskal–Wallis
- Resampling techniques
- Statistical analysis and data visualization in R or Python

References

[1] W.N. Venables, D.M. Smith, others, An Introduction to R, CiteSeer, 2009.

[2] G. Grolemund, H. Wickham, R for data science, 2017. <https://r4ds.had.co.nz>.

[3] R Programming Courses, n.d. <https://www.edx.org/learn/r-programming>.

[4] I.C.R. Group, Intersalt: An international study of electrolyte excretion and blood pressure. Results for 24 hour urinary sodium and potassium excretion, Brit. Med. J. (1988) 319–328.

[5] Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline, third ed., vol. 28, CLSI, 2010.

[6] S.S. Shapiro, M.B. Wilk, An analysis of variance test for normality (complete samples), Biometrika. 52 (1965) 591–611.

[7] S.S. Shapiro, R. Francia, An approximate analysis of variance test for normality, J. Am. Stat. Assoc. 67 (1972) 215–216.

[8] H.W. Lilliefors, On the Kolmogorov–Smirnov test for normality with mean and variance unknown, J. Am. Stat. Assoc. 62 (1967) 399–402.

[9] T.W. Anderson, D.A. Darling, A test of goodness of fit, J. Am. Stat. Assoc. 49 (1954) 765–769.

[10] A.S. et mult. al., DescTools: tools for descriptive statistics, 2019. <https://cran.r-project.org/package = DescTools>.

[11] F.E. Grubbs, Procedures for detecting outlying observations in samples, Technometrics. 11 (1969) 1–21.

[12] R.B. Dean, W. Dixon, Simplified statistics for small numbers of observations, Anal. Chem. 23 (1951) 636–638.

[13] D.B. Rorabacher, Statistical treatment for rejection of deviant values: critical values of dixon's "Q" parameter and related

subrange ratios at the 95% confidence level, Anal. Chem. 63 (1991) 139−146.
[14] J.W. Tukey, Exploratory Data Analysis, Reading, MA, 1977.
[15] W. Chavuenet, A Manual of Spherical and Practical Astronomy, 1871.
[16] F.J. Anscombe, Rejection of outliers, Technometrics. 2 (1960) 123−146.
[17] D. Dalpiaz, Applied statistics with R, 2018. <https://daviddalpiaz.github.io/appliedstats/applied_statistics.pdf>.
[18] G. Kerns, Introduction to probability and statistics using R, 2018. <https://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf>.
[19] G.E. Box, D.R. Cox, An analysis of transformations, J. R. Stat. Soc. Series B. Stat. Methodol. 26 (1964) 211−243.
[20] R.A. Peterson, Using the best normalize package, 2018. <https://cran.r-project.org/web/packages/bestNormalize/vignettes/bestNormalize.html>.
[21] W.N. Venables, B.D. Ripley, Modern Applied Statistics With S, fourth ed, Springer, New York, 2002. <http://www.stats.ox.ac.uk/pub/MASS4>.
[22] W.E. Deming, Statistical Adjustment of Data, 1985 ed, Dover Publications, 1943.
[23] K. Linnet, Evaluation of regression procedures for methods comparison studies, Clin. Chem 39 (1993) 424−432.
[24] F.M. Ekaterina Manuilova Andre Schuetzenmeister, mcr: method comparison regression, 2014. <https://CRAN.R-project.org/package = mcr>.
[25] T. Therneau, Deming: Deming, Theil−Sen, Passing−Babloc and total least squares regression, 2018. <https://CRAN.R-project.org/package = deming>.
[26] H. Passing, W. Bablok, A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, part I, Clin. Chem. Lab. Med. 21 (1983) 709−720.
[27] H. Passing, W. Bablok, Comparison of several regression procedures for method comparison studies and determination of sample sizes application of linear regression procedures for method comparison studies in clinical chemistry, part II, Clin. Chem. Lab. Med. 22 (1984) 431−445.
[28] W. Bablok, H. Passing, R. Bender, B. Schneider, A general regression procedure for method transformation. Application of linear regression procedures for method comparison studies in clinical chemistry, part III, Clin. Chem. Lab. Med. 26 (1988) 783−790.
[29] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, et al., PROC: an open-source package for R and S + to analyze and compare ROC curves, BMC Bioinformatics. 12 (2011) 77.

Self-assessment questions

1. The SD is:
   a. the average difference between any value and the mean
   b. the variance divided by 2
   c. the square root of the variance
   d. the mean squared difference between individual values and the mean
   e. the mean difference between any value and the mean
2. If the probability of an event occurring is 90%, what is the corresponding odds?
   a. 9
   b. 10%
   c. 90
   d. 10:1
3. Regarding Passing–Bablok regression, which of the following is false?
   a. It is computationally expensive.
   b. It is a form of so-called robust regression.
   c. It has unique popularity in the field of clinical chemistry.
   d. It relies on the principle of minimization of squared residuals.
   e. It may give different slope and intercept if the units of reporting are changed.
4. If a linear regression line has a slope of 1.07 and *y*-intercept of 12:
   a. There is a proportional bias of 12.
   b. There is a constant bias of 12.
   c. There is a constant bias of 0.07.
   d. The correlation coefficient will be negative.
   e. *y* values will typically be less than the corresponding *x* values.
5. In a population of 190 subjects with a disease prevalence of 10%, a diagnostic test is performed. It is found that there are 15 true positives, 4 false negatives, 160 true negatives, and 11 false positives. The posttest odds of a positive test is:
   a. 57%
   b. 1.36
   c. 0.11
   d. 78%
6. The normal QQ plot is:
   a. useful to determine if the data follow a χ^2 distribution
   b. a plot of the quantiles of the sample distribution against the theoretical quantiles of the normal distribution
   c. sigmoidal when the sample distribution is Gaussian
   d. a way of comparing the mean of two distributions
7. Which of the following is true of OLS regression:
   a. It assumes that there is no error in the *x*-axis data.
   b. It assumes that the error in the *y*-axis data is homoscedastic.
   c. Its residuals are vertical.
   d. The procedure minimizes the sum of the squared residuals.
   e. All of the above.
8. In a distribution that has positive skewness:
   a. The mean and the median are likely equal.
   b. The mean and the mode are likely equal.
   c. The median is likely lower than the mean.
   d. The median is likely higher than the mean.
   e. The distribution is platykurtic.
9. $1/x$ weighting is useful in linear regression when:
   a. There is homoscedastic error.
   b. The error is constant.
   c. The scatter in the data shows no spreading as *x* increases.
   d. There is a need to make the small values of *x* affect the regression line more than the large values of *x*.
   e. Deming regression is not available in your spreadsheet program.
10. Which of the following cannot be used to assess the normality of a distribution?
    a. a histogram
    b. a QQ plot
    c. a Shapiro–Wilk test
    d. a Kaplan–Meier curve
    e. calculation of the skewness and kurtosis

Answers

1. c
2. b
3. d
4. b
5. b
6. b
7. e
8. c
9. d
10. d

Chapter 3

Reference intervals: theory and practice

Victoria Higgins, Michelle Nieuwesteeg, and Khosrow Adeli
Clinical Biochemistry, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada

Learning objectives

After completing this chapter, the reader should be able to:

- Discuss the value of reference intervals to laboratory medicine.
- Describe the processes of establishing reference intervals, including selection of a reference population, consideration of analytical and preanalytical variables, and statistical methods for calculating reference intervals.
- Identify the challenges associated with establishing reference intervals for subgroups of the population (i.e., pediatrics and geriatrics).
- Explain the process, value, and limitations of transferring and verifying reference intervals.
- Understand the concept, challenges, and limitations of harmonization in laboratory medicine, especially reference intervals.

Introduction

Concept of reference intervals and importance to laboratory medicine

Reference intervals are fundamental tools used by medical practitioners to interpret patient laboratory test results and help differentiate between healthy and unhealthy individuals [1]. Sometimes referred to as "normal" or "expected" values, reference intervals provide the range of laboratory test results that would be expected in a healthy population. Therefore reference intervals serve as "health-associated benchmarks," where results that fall outside of the reference interval may be interpreted as abnormal, indicating the need for additional medical follow-up and/or treatment [2].

Clinical laboratories provide valuable information to physicians and health care providers to aid in clinical decision-making, including diagnosis and management of disease [3,4]. Over the past several decades, there have been significant technological advancements in laboratory medicine, leading to increased accuracy and precision of biochemical assays, improved analytical quality of laboratory procedures, and a growing number of novel biomarkers that are assessed to aid in disease diagnosis, prognosis, and monitoring [5]. However, the tools used to interpret laboratory test results and their impact on patient care and safety have received less attention from laboratorians and clinicians. Arguably, the quality of the reference intervals used to interpret test results may be equally important as the quality of the result itself [6]. In modern medicine, it is essential that reference intervals are updated frequently to keep pace with changing methodologies and that well-defined reference intervals are established for novel tests. It is clear that the lack or inappropriate use of reference intervals can result in misdiagnosis, unnecessary medical follow-up, increased patient risk, and higher health care costs. Although the concept and use of reference intervals are generally straightforward, establishment of accurate reference intervals can be complex, requiring recruitment of healthy subjects, awareness of analytical and preanalytical factors that can affect testing, careful statistical analysis, and consideration of important covariates (e.g., age, sex, and ethnicity). Establishing a new reference interval requires a minimum of 120 healthy reference subjects per partition, as recommended by the Clinical and Laboratory Standards Institute (CLSI) [7]. Given the complexity of establishing a new reference interval, most laboratories adopt reference intervals provided by the instrument manufacturer or attempt to transfer and validate an existing reference interval. Verification of a reference interval requires a much smaller sample size of healthy subjects (i.e., minimum $n = 20$ per partition [7]), and is thus much more feasible for individual laboratories to perform. This chapter will review procedures for establishing reference intervals, factors that affect reference interval determination, implementation and transference of reference intervals, global reference interval initiatives, and approaches for reference interval harmonization (hRI).

Reference intervals versus clinical decision limits

Statistically, reference intervals are defined as limiting values within which a specified percentage (usually the central 95%) of values from an apparently healthy reference

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00003-X

population would fall. According to this definition, typically, the 2.5th and 97.5th centiles of the test result distribution are used to define the upper and lower reference limits, as depicted in Fig. 3.1. This means that 2.5% of individuals with the highest test results and 2.5% of individuals with the lowest test results are excluded (reviewed in Ref. [8]). In some cases, this definition may be modified. For example, only the upper or lower limit may be considered for analytes, where only one side of the distribution is clinically significant. In addition, for some analytes, a different centile may be used to provide greater specificity. For instance, the 99th centile of a healthy population is used to establish the upper reference limit for cardiac troponin [9].

The International Organization for Standardization (ISO) directive for medical laboratories (ISO 15189) requires biological reference intervals or clinical decision limits to be available and appropriately reviewed and updated [10]. Clinical laboratories use reference intervals to interpret most laboratory test results; however, decision limits are used for a limited number of biomarkers. Reference intervals describe the distribution of values from an apparently healthy population. Individual test results can be compared to the reference interval to determine if the test value fits within the defined reference range, established from the distribution of values based on an apparently healthy reference population. In contrast, clinical decision limits, as the name suggests, distinguish between particular clinical states. Decision limits provide threshold values; values exceeding or falling below the threshold indicate a significantly higher risk of a clinical outcome or may satisfy criteria for diagnosis of a specific disease [10]. Furthermore, multiple thresholds may be used to indicate levels of increasing risk for a disease (e.g., borderline high and high). In contrast to reference intervals, decision limits are often established from clinical outcome studies, including prospective cohort studies and meta-analyses, but can also be based on consensus of expert panels. The CLSI EP28-A3c guidelines state that "when decision limits determined by national or worldwide consensus exist, these limits, rather than reference intervals, should be reported" [7].

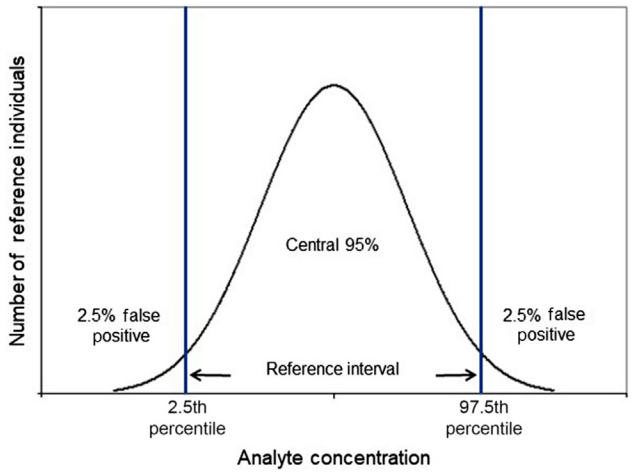


FIGURE 3.1 Graphical representation of a reference interval. A reference interval is defined as the central 95% of laboratory test results from a healthy, reference population. An example of a normal distribution for analyte concentration is shown.

Current gaps and recent initiatives in reference interval establishment

Patients and their health care workers assume that reference intervals are readily available to medical laboratories and health care institutions performing clinical tests and interpreting test results. Unfortunately, this is not always the case and many laboratories depend on outdated and incomplete information when reporting test results. In many cases, the same reference intervals are used to interpret test results for children and adults, which can lead to erroneous and inaccurate interpretation. Available reference intervals also suffer from several limitations including small sample sizes, and the use of outpatient or hospitalized patient samples. There has thus been an urgent need for the development of new up-to-date laboratory reference interval databases based on healthy children, adult, and geriatric populations. It is often difficult to obtain sufficient numbers of healthy reference samples to stratify reference intervals according to multiple covariates [11]. This can pose a major challenge when establishing accurate reference intervals, particularly for pediatric or geriatric populations where it is more difficult to collect large numbers of blood samples from healthy individuals. As a result, many existing reference intervals were determined decades ago using methods and/or instruments that are no longer relevant given the technological advances in laboratory testing. Further, reference intervals are lacking for many novel/emerging biomarkers. Thus current and future initiatives should consider these gaps in the design of their reference interval studies.

A number of international initiatives are recognizing the current gaps in reference intervals and the need to establish values that are up-to-date and appropriate for the local population. In Canada, Canadian Laboratory Initiative on Pediatric Reference Intervals (CALIPER) has partnered with the Canadian Health Measures Survey (CHMS) to evaluate data collected from over 12,000 Canadians, ranging in age from 3–79 years, in order to establish reference intervals for over 50 common biochemical, endocrine, and hematology biomarkers for pediatric, adult, and geriatric populations [12–14]. CALIPER also has established a comprehensive pediatric reference interval database of age- and sex-stratified reference intervals for over 170 pediatric biomarkers currently available (described further below) [15–31]. In Australia, a direct reference interval study, called Aussie Normals, has surveyed 1876 healthy male and female participants aged 18–65 years.

All individuals were from the Australian Capital Territory, which has a multicultural population, representative of other regions in Australia. The study established reference intervals for 91 analytes, with individuals excluded on the basis of pregnancy, diabetes, renal disease, and cardiovascular disease [32]. In Europe, several reference interval initiatives have been established. Nordic Reference Interval Project (NORIP) is a collaborative initiative among five Nordic countries to measure 25 of the most common chemistry analytes [33]. Laboratory Reference Ranges for Turkey, conducted by the Association of Clinical Biochemists in Turkey, is a reference interval initiative aimed at updating all outdated reference intervals for the Turkish population [11]. REALAB, an Italian reference interval initiative, has developed reference intervals for 23 chemistry analytes plus 13 additional markers, based on 61,246 samples collected from hospital patients [34]. Although this is not the preferred method recommended by the CLSI, many reference studies use this approach to estimate reference intervals due to the challenges of collecting large numbers of blood samples from healthy individuals.

Methodological approaches to establishment of reference intervals

The CLSI and the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) provide detailed guidelines for the clinical laboratory community on defining, establishing, and verifying reference intervals [7,35–41]. The recommendations for the selection of reference individuals, preanalytical and analytical considerations, outlier exclusion, and statistical methods for reference interval determination and partitioning are described below and are summarized in Fig. 3.2.

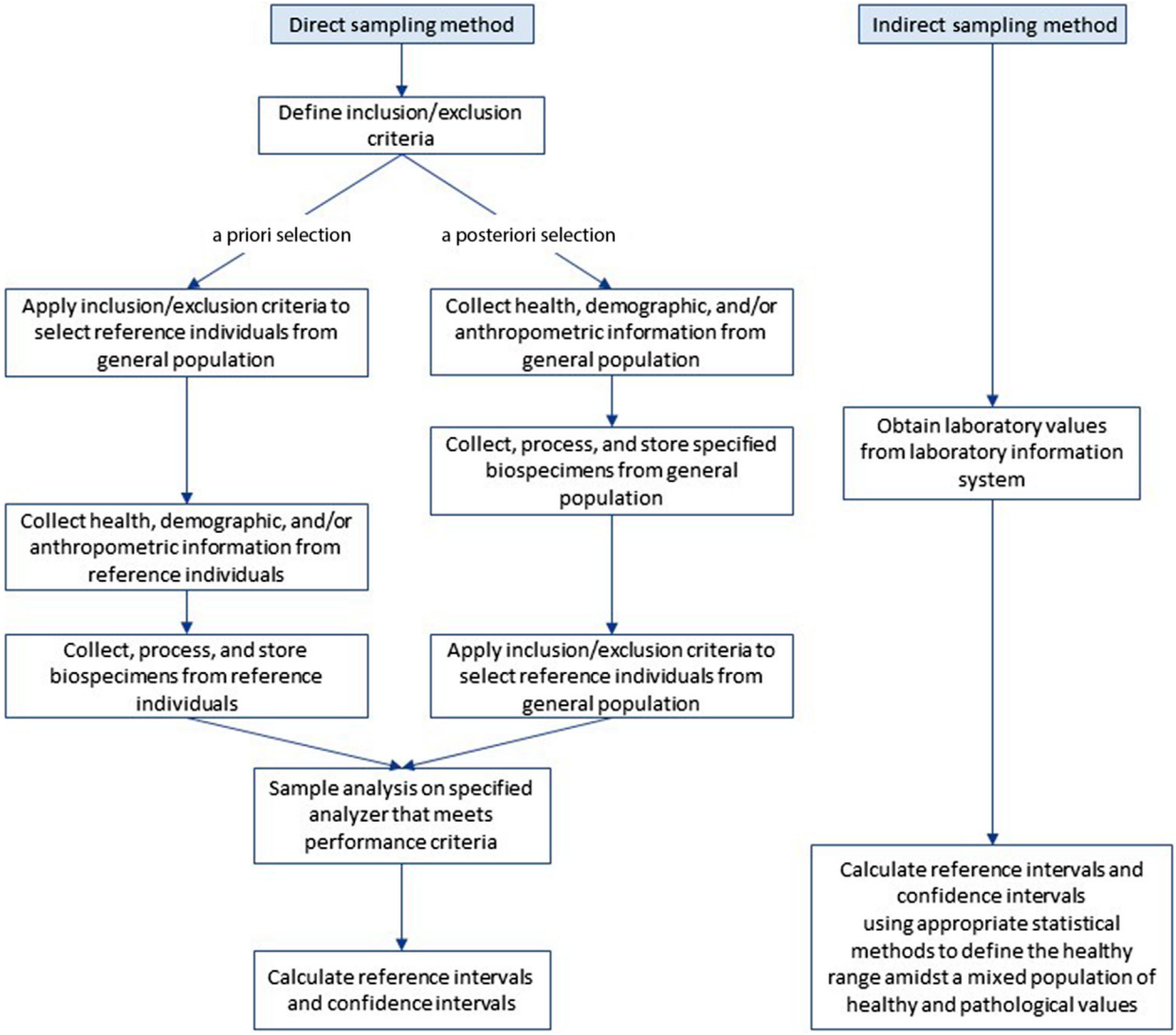


FIGURE 3.2 Flowchart of direct and indirect reference interval establishment processes.

Population-based reference intervals

Population-based reference intervals are derived from a group of well-defined reference individuals that, ideally, are similar to the target patient in all respects other than the disease condition under investigation. Population-based reference intervals are widely used in contrast to individual reference intervals, based on values obtained from the same individual, as they are more impractical to establish and implement. Thus this chapter will refer only to the establishment of population-based reference intervals.

Selection of reference individuals

The quality of a reference interval can depend heavily on the selection and recruitment of appropriate reference individuals [8,11,42]. According to the CLSI EP28-A3c guidelines, reference intervals should be based on direct sampling of healthy individuals from the age group(s) of interest, using well-defined inclusion/exclusion criteria. Important factors to consider when selecting reference individuals include *direct* versus *indirect sampling* and application of a priori versus a posteriori selection criteria.

Direct versus indirect sampling

Direct sampling is the preferred method recommended by the CLSI [7] and involves the selection of healthy individuals using defined inclusion/exclusion criteria. Using this method, "unhealthy" individuals are not included, and thus do not jeopardize the validity of the reference interval. The challenges associated with the direct sampling method, however, include the difficulty and cost associated with collecting a sufficient number of blood samples from healthy individuals to calculate accurately reference intervals [39]. Statistical methods used to establish reference intervals using data obtained by direct sampling are discussed in the section "Statistical determination of reference intervals."

In contrast, *indirect sampling* uses laboratory values obtained from a patient population, which includes both healthy and pathological subjects, to estimate reference intervals. This approach is easier and less costly than the direct sampling technique. However, the use of indirect sampling is not recommended by the CLSI due to the risk of including values from diseased individuals, which may result in skewed, broader, or less sensitive reference intervals [7]. However, the CLSI does acknowledge that, in some cases, the use of indirect sampling may be necessary when it is too difficult to recruit sufficient numbers of healthy individuals [7]. This is often the case in pediatrics, where issues such as small sample volume and parental consent can pose significant challenges to recruitment of a healthy reference population. Statistical methods that have been proposed to calculate reference intervals based on the indirect sampling method include the Hoffmann [43], Bhattacharya [44], modified Bhattacharya to accommodate Gamma distributions [45], and methods proposed by Arzideh et al. [46,47] for data that typically require a Box−Cox transformation and Gaussian distributed data. All proposed methods attempt to define the range of data explained by healthy subjects amidst a mixed population of both healthy and pathological samples. Furthermore, all methods assume the majority of observations in the data set are from healthy individuals, an assumption difficult to ascertain depending on the patient population employed.

A priori versus a posteriori sampling

Both a priori and a posteriori sampling approaches can be used to establish reliable reference values. The determination of which method to use often depends on the nature of the study. In a priori sampling, well-defined inclusion/exclusion and partitioning criteria must be established before reference individuals are selected. In addition, an appropriate questionnaire designed to reveal these criteria is necessary. Exclusion criteria should exclude individuals from the reference sample group based on assessments indicating lack of good health (e.g., recent or chronic illness, use of prescription or nonprescription medication, use of tobacco or alcohol, and diet) [7]. Reference intervals can be partitioned by several covariates that are known to affect the level of the specific analyte (i.e., sex, age, Tanner stage, and ethnicity), and therefore the questionnaire should include these factors as well. This procedure is typically applied for analytes with well-established methods and laboratory procedures, and known sources of biological variation. Information from the literature is used to develop exclusion and partitioning criteria, which are included in the questionnaire to exclude individuals from the sampling process [7]. In contrast, in the a posteriori approach, exclusion and partitioning occur after sample collection and analysis. This approach is recommended for analytes that are new or for which well-documented information does not exist in the literature regarding laboratory procedures and biological variation. However, this approach may require a more detailed questionnaire to capture information on all potential covariates that may be used as exclusion or partitioning criteria [7].

Preanalytical and analytical variables

A number of preanalytical and analytical variables can influence test results and reference intervals. Therefore these variables must be carefully considered and controlled in both the patient population and in reference individuals. The CLSI EP28-A3c guideline lists a variety of preanalytical and analytical variables that should be considered when developing reference intervals [7].

Preanalytical variables include subject preparation, specimen collection, and specimen handling considerations. Depending on the analyte, subjects may need to be

in a fasted or a nonfasted state, may require a rest period from physical activity prior to sample collection, or may need to abstain from substances that could affect analyte levels, such as pharmacological agents, caffeine, alcohol, tobacco, or vitamin C. In addition, some analytes vary with circadian rhythm or stage of the menstrual cycle (in the case of females) and, thus, these factors should also be considered. Other subject-related variables may include prior diet and stress level. Specimen collection considerations include time of day, environmental conditions during collection, body posture, specimen type, blood flow, equipment, tourniquet time, collection technique, sample volume, and anticoagulants. Regarding specimen handling, it is important to consider variables such as transport, storage, clotting, separation of serum/plasma, and sample preparation for analysis.

Analytic variability of the method used for measurement can also impact reference intervals, and thus it is important that the measurement of reference samples and patient samples are performed in an identical manner [2,7]. The methods used should be clearly described, and between-run analytical imprecision, limit of detection, linearity, recovery, reportable range, interference characteristics (i.e., hemolysis, icterus, and lipemia), and traceability should be reported. Instrument and equipment maintenance and normal operation must also be equivalent, as should quality control procedures for reagents, calibration, controls, and calculation methods. Additional consideration should include reagent lot-to-lot variability, instrument-to-instrument variability (if multiple analyzers are used), and technologist variability [7].

Outlier exclusion

Before reference intervals are calculated, the data should first be examined for the presence of observations with extreme values that have the potential to influence the estimation of reference values. Although reference populations are usually selected to minimize the number of outliers in the distribution, in actuality, this is often impossible to achieve. Thus the frequency distribution of the data should first be inspected for the presence of outliers. Frequency histograms or box plots can be used to inspect visually the data for the presence of outliers. However, statistical methods should be applied to remove outliers prior to reference interval determination. According the CLSI EP28-A3c guidelines, two methods are recommended for outlier removal: Dixon's test or Tukey's method [7].

Dixon's test calculates the absolute difference (D) between the suspected outlier (which can be an extreme large or extreme small value) compared with the next largest/smallest observation and compares this with the range of all observations (R), including extremes. In Dixon's test, a ratio of 1/3 is used as a cutoff, where, if D/R exceeds the 1/3 ratio (meaning that D is greater than or equal to 1/3 of the range R), then the outlier should be removed. If two or more outliers are suspected on the same side of the distribution, the 1/3 rule can be applied to the least extreme outlier and removed collectively with more extreme values [48,49]. Dixon's test requires the data be normally distributed, and data that do not follow a Gaussian distribution should be transformed before applying Dixon's test. In addition, it should be noted that Dixon's test should only be applied once to remove outliers, which may not always be practical depending on the nature of the distribution. Furthermore, Dixon's tests preferentially identify and remove outliers with high values compared with those at the lower end of the normal distribution, which should be considered before choosing to use Dixon's test or Tukey's method for outlier removal [48].

Tukey's method involves calculating the 25th and 75th percentiles of the data set ($Q1$ and $Q3$, respectively), and subsequently using these values to determine the interquartile range (IQR; $Q3-Q1$). The IQR is used to determine the boundaries for outlier exclusion, where the lower boundary is calculated by $Q1-1.5\times\text{IQR}$ and the upper boundary is calculated by $Q3+1.5\times\text{IQR}$. Any value below the lower boundary or above the upper boundary should be removed as outliers [50]. Similar to Dixon's test, Tukey's method requires a normally distributed data set. The data may be transformed to achieve normality using a method such as Box–Cox transformation, or an adjusted Tukey test may be used on a skewed data set [51]. Unlike Dixon's test, however, the Tukey method may be applied to a data set more than once to remove outliers and does not preferentially favor removal of outliers from one side of the distribution over the other. Thus Tukey's method may be better suited for outlier removal when both the upper and lower limits of the reference interval must be determined [2].

Partitioning and minimum sample size

Partitioning is used when significant differences exist in analyte levels between subgroups, such as between sexes or different age groups. Partitioning or establishment of separate reference intervals for different subgroups of a reference population may be necessary for certain analytes. For established analytes, the clinical or physiological significance of partitioning may already be known. For new markers, however, this information may not be available. The CLSI recommends that for each partition, or subgroup, a minimum number of 120 reference individuals should be included in the calculation of the reference interval [7].

The decision to partition reference intervals into distinct subgroups is often based first on visual inspection of the reference distribution. For instance, bimodal or polymodal

distributions contain more than one peak and may be an indication that the data contain two or more distributions, which should be partitioned according to age, sex, and/or other factors. The final decision to partition, however, should be based on a statistical test to determine whether the difference in the means between two subgroups is statistically significant. There are several tests that can be used for the purpose of partitioning. Sinton et al. [52] recommend that partitioning is only required if the difference in means is greater than 25% of the 95% reference interval. In contrast, Harris and Boyd [53] suggest partitioning if the ratio of the subgroup standard deviations is ≥ 1.5. It should be noted, however, that the Harris and Boyd method requires Gaussian distribution of the data. The CLSI EP28-A3c guidelines also list an alternate method proposed by Lahti et al. that can be used on data of any distribution and focuses on distances between the reference limits instead of the distances between the means [7,54]. While partitioning of categorical variables, such as sex, is relatively straightforward using the methods described above, partitioning based on continuous variables, such as age, can be more complex. In these cases, continuous reference intervals may better reflect the relationship between the age and the analyte concentration [55–57].

Statistical determination of reference intervals

The CLSI EP28-A3c guideline primarily focuses on two statistical methods to calculate reference intervals: the traditional nonparametric method, which requires no assumption about the distribution of the data, and the robust method, which requires fewer observations to calculate the reference interval and is not sensitive to outliers [7]. Parametric methods may also be used to calculate 95% reference intervals, although this method assumes the data follow a Gaussian distribution. Various statistical approaches for calculating reference intervals based on the direct sampling method will be discussed in this section.

Nonparametric analysis

The traditional nonparametric rank order analysis is recommended by the CLSI to calculate reference intervals due to the simplicity of this statistical method and the fact that this method makes no assumptions regarding data distribution given the common non-Gaussian distribution of analyte measurement. In this approach, the reference values are sorted from the lowest to highest and rank-ordered, with the smallest value receiving a rank of 1 and the largest value receiving a rank of n [7]. The lower reference limit is then determined by calculating the observation that corresponds to the 2.5th percentile $[r_1 = 0.025(n + 1)]$ and the 97.5th percentile $[r_2 = 0.975(n + 1)]$. If the values of r_1 and r_2 are not integers, then one should interpolate between the two limiting values. It should be noted that the nonparametric rank method is not recommended for data with less than 120 reference values. The CLSI also recommends that 90% confidence intervals are calculated for each end point of the reference interval [7]. An alternative nonparametric method is the Harrell–Davis nonparametric bootstrap method [58], which provides a more accurate estimation of the reference interval by continuous resampling, but requires more complex computation and demands more statistical expertise.

Robust analysis

The CLSI recognizes that, in some cases, it is not possible to obtain a minimum number of 120 reference samples. In this case, the robust method developed by Horn and Pesce [1,59] is recommended as an alternative approach, although the method is more complicated. The robust method assigns each reference value a statistically calculated "weight." The weights are proportional to each value's distance from the mean, with greater weights assigned to central values than the more distant values. The weighted values represent a new data set with a distribution that closely approximates the underlying distribution of the data. Bootstrapping is used to repeat the algorithm multiple times, generating successive distributions, which each time resemble the underlying distribution more closely. After sufficient bootstrapping, an estimate of the 95% reference interval can be computed. As few as 20 reference samples may be used to estimate reference intervals using the robust method, however, in order to achieve 90% confidence level, the minimum recommended number of observations is 80 [7].

Parametric analysis

For data with a Gaussian distribution, reference intervals can be calculated using parametric analysis, which uses the mean and 1.96 standard deviations of the data set to calculate the 2.5th and 97.5th percentiles. Prior to parametric analysis, it is important that the data are checked using statistical tests (such as the Anderson–Darling test recommended by the IFCC) to ensure that the data fit a Gaussian distribution. If the data are skewed, transformation can be used to approximate a Gaussian distribution. However, the IFCC recommends a two-step approach for data transformation to correct for skewness and kurtosis, respectively, such as the method outlined by Harris and DeMets [60]. When the data are Gaussian or can be transformed to be Gaussian, the parametric method produces reference intervals with the least biased estimates (i.e., most accurate) and the lowest mean squared error (i.e., most precise) across a wide range of sample sizes (i.e., $n = 40–480$), according to a systematic comparison of the performance of traditional

nonparametric, parametric, and robust methods [61]. As the distribution of reference values for most analytes is often not Gaussian, parametric analysis often requires complex statistical theory and computer programs to test whether transformed data conform to a Gaussian distribution. Due to the complexities of parametric analysis and the common non-Gaussian distribution of analyte concentration, the nonparametric approach is recommended by the CLSI [7].

Covariates that affect reference interval determination

The specificity and sensitivity of reference intervals may be improved by partitioning the reference population where appropriate. A number of physiological factors may contribute to clinical variation in analyte levels within a population, which should be considered when establishing reference intervals [62]. Reference intervals are frequently partitioned by age and sex. However, other important covariates may also influence analyte concentration including pregnancy, organ maturity, growth, hormonal changes, and sexual development. These covariates are not relevant to all age groups or analytes, but may result in important differences in some cases. For instance, levels of fertility hormones may be different in pregnant versus nonpregnant women, requiring separate reference interval partitions. In addition, during development, levels of fertility hormones, important for indicating pediatric endocrinopathies, may vary with age, sex, and sexual development (Tanner stage), and should be partitioned accordingly.

Ethnicity is also emerging as a recognized factor that contributes to the differences in the concentration of some analytes, as a result of genetic and/or environmental factors [1,63,64]. Currently, the majority of reference intervals are based on a predominant Caucasian population, which could be inappropriate for the ethnically diverse patients served by many laboratories. Ethnic differences in analyte measurements (e.g., routine chemistry, fertility, endocrine, cancer, and hematologic markers) between several ethnic groups (e.g., Black, Caucasian, East Asian, Hispanic, South Asian, South East Asian, and West Asian) were recently reviewed, highlighting potential differences between the ethnicities for a subset of analytes [65]. Thus establishment of ethnic-specific reference intervals for a subset of biomarkers may lead to more accurate diagnosis/treatment and improved patient care. However, prior to implementing ethnic-specific reference intervals in clinical practice, it is important to consider concurrently reported significant ethnic differences, population demographics, environmental factors, and disease risks to ensure these reference intervals will improve clinical assessment.

Partitioning by genetic differences provides another potential means of reducing interindividual variation within a reference interval, improving their clinical utility. Including this additional covariate may reduce misidentification of unusual test results that are caused by genetic variation [66]. A statistical framework to estimate the utility of partitioning reference intervals based on genetic variation has been proposed [66]. This exercise uncovered important considerations when deciding whether to partition based on genetic variation, including the size of the genetic effect and the prevalence of the polymorphism in the population [66]. Furthermore, the complexity of genetic variation effect on reference values would have to be simplified inherently to permit reference interval partitioning, as multiple genes and environmental factors contribute to the total variance in analyte concentration [66]. Genetic information has been considered during the establishment of reference intervals for apolipoprotein E [67] and haptoglobin [68]. Incorporating genetic variation information into reference interval establishment may become common with clinically available whole-genome data and a better understanding of the effect of genetic polymorphisms on laboratory data [69].

Age-dependent reference curves

The statistical methods most commonly employed for reference interval calculation provide discrete upper and lower reference limits for a particular age and sex subgroup. However, biomarker concentrations do not abruptly change with age, as discrete reference intervals suggest. Rather, biomarker concentration changes continuously with age, particularly throughout the pediatric age range due to rapid periods of growth and development. Therefore rather than treating age as a categorical variable, it can be treated as a continuous variable in the development of age-dependent reference curves. Reference curves have been widely used to provide accurate reference data for body measures in the pediatric population, including growth reference charts by the World Health Organization [70] and the Centers for Disease Control and Prevention [71]. Continuous reference intervals are now emerging for laboratory biomarkers, using both direct [72,73] and indirect sampling methods [74–78]. Although indirect sampling methods are more feasible, direct sampling methods are still preferred to establish reference intervals due to reduced preanalytical and analytical variations and minimization of errors [79]. Several statistical methods for direct reference interval establishment have been proposed including parametric (e.g., fractional polynomials [80]), semiparametric (e.g., lambda-mu-sigma method [81]), and nonparametric methods (e.g., nonparametric quantile regression [82,83]). Statistical methods for developing continuous reference intervals using the indirect sampling method have also been proposed, including a method developed by Zierk et al. [75,78]. Although continuous reference intervals better reflect the dynamic trend in biomarker concentration with age, their implementation in clinical practice

remains an obstacle, as laboratory information systems are currently unable to accommodate a mathematical function for laboratory test interpretation. Therefore while continuous percentile curves provide a more accurate representation of normative biomarker concentration, age and/or sex partitions are still required for practical use, which inherently reduces the complexity of true age-related trends.

Pediatric and geriatric reference intervals: challenges and recent advances

While a number of ongoing national initiatives have begun to close the gaps in establishing up-to-date reference intervals, many of these studies have focused on the adult population, generally analyzing samples collected from reference individuals between 18 and 65 years of age. Unfortunately, reference intervals for older and younger individuals pose unique challenges to sample collection and reference interval calculation, often requiring stratification by multiple covariates and an increased number of reference individuals. Thus large knowledge gaps remain in reference intervals for pediatric and geriatric populations. The following sections will describe the challenges and procedures for establishing reference intervals in these age groups, and will review recent initiatives aimed at addressing these gaps.

Pediatric reference intervals

Establishing reference intervals specific for the pediatric population is a task of utmost importance but also of great complexity, requiring both time and resources. Traditionally, pediatric laboratory tests have been interpreted based on adult normative values. However, children cannot be viewed as simply "small adults" in medical practice. Children and adolescents differ from adults in many aspects, including physical size, organ maturity, rates of growth, developmental stages, and immune and hormone responsiveness [2]. In addition, newborns respond to infections differently from adults, as they have little to no history of infectious disease exposure and often require special laboratory testing [5]. Furthermore, puberty is a unique developmental stage that occurs in adolescence, which can profoundly affect various analyte levels. For example, levels of sex and growth hormones are low in infants, but increase during and after puberty, while markers of bone growth (e.g., alkaline phosphatase) are elevated in children and then decrease in adulthood with the cessation of bone growth [12]. These notable trends highlight the importance of developing pediatric-specific reference intervals. A major challenge faced during reference interval determination in the pediatric population is recruiting a sufficiently large and healthy reference population. There is a need for large-scale public relation campaigns to recruit a sufficient number of participants, which can be costly and time-consuming. Parental consent is needed prior to collecting blood from young subjects and thus poses an additional barrier. Also a sufficient sample volume can be difficult to be obtained from children and certified phlebotomists with special training are required to perform pediatric blood collection.

Pediatric reference intervals should reflect the dramatic physiological and hormonal changes that take place over the course of development. Body weight and length increase at a considerable rate during the first year of life. During puberty, important hormonal changes take place in addition to accelerated growth and sexual maturation. These events are all unique to the pediatric age range and can dramatically affect the levels of many analytes. It is, therefore, incredibly important that pediatric reference intervals reflect the differences in development and physiological function at different ages. Extensive partitioning is required to ensure reference intervals are specific for these different physiological stages. Additional complexity arises from subjectivity in differentiating between statistically relevant and clinically relevant partitions. Sometimes partitions will be deemed statistically relevant (e.g., according to the Harris and Boyd method), but clinically, it may not be necessary to have separate partitions for specific age and/or sex categories. Alternatively, partitions may not be statistically relevant, but clinically, a known physiological change may justify separating specific age and/or sex partitions as appropriate.

Regional initiatives to address the gaps in pediatric reference intervals have been performed with an extensive number of healthy children across the pediatric age range and for both sexes. These include the KiGGS (German Health Interview and Examination Survey for Children and Adolescents) study in Germany, the CALIPER initiative in Canada, the Children's Health Improvement through Laboratory Diagnostics (CHILDx) Study in the United States, and initiatives of the Scandinavian Societies of Clinical Chemistry. The KiGGS study, part of the health monitoring system of the Robert Koch Institute, aimed to collect, for the first time, comprehensive and nationwide data on the health status of children and adolescents aged 0–17 years. From 2003–06, information was collected by means of surveys, medical examinations, tests, and laboratory analyses from 17,641 children at 167 locations in Germany. The KiGGS study successfully addressed their aim of collecting comprehensive, nationwide data on the health status of German children and adolescents, evident through several publications [84,85]. These include studies on general health (i.e., mental health issues, subjective health, and accidental injuries), chronic diseases (i.e., bronchial asthma, hay fever, and neurodermatitis), and health factors (i.e., alcohol consumption, physical activity, sports, and tobacco consumption). More recently, this research group from Germany developed percentile charts from laboratory data collected from 10 German centers, including

laboratory results from over 350,000 pediatric patients. They also created a website application (www.pedref.org/hematology) that allows visualization of hematology test results using percentile charts and z-scores, providing precise test result interpretation. CALIPER is a Canadian initiative that began in 2009 to address the critical gaps in pediatric reference intervals. Since its inception, CALIPER has recruited over 10,000 children and adolescents aged 0− < 19 years and has established age- and sex-specific reference intervals for over 170 biomarkers, including routine chemistry analytes, proteins, enzymes, fertility hormones, and tumor markers [15−20,23−25,28,29,31]. In 2002 the CHILDx was initiated at ARUP Laboratories to determine pediatric reference intervals for a number of clinical laboratory assays using samples collected at ARUP and Primary Children's Medical Center in Salt Lake City, Utah. Reference intervals have been established for 35 assays (goal of 66 assays) for children aged 6 months to 6 years and 58 assays (goal of 144 assays) for children aged 7−17 years [86−93]. The Nordic countries have also been successful in establishing pediatric reference intervals for 21 biochemical properties. These reference intervals were established using pediatric samples (aged 5−19 years) collected from schools in the Copenhagen area in Denmark as part of THE COPENHAGEN Puberty Study (ClinicalTrials.gov ID: NCT01411527) conducted in 2006−08 [94].

Geriatric reference intervals

Another subpopulation that poses unique challenges to establishing reliable reference intervals is the geriatric population. The geriatric population seeks health care more often than individuals representing the age group 20−50 years [95]. However, there are relatively few published reference interval studies specific for the geriatric population and statistical approaches for their calculation vary widely [96]. Defining and obtaining a healthy reference population is essential for calculating representative reference intervals, and this is one of the most challenging tasks when studying the geriatric population. As individuals age, there is a large increase in the proportion of individuals who are not healthy. In 2013 Horn and Pesce reported that to obtain healthy reference samples from individuals aged 70−80 years old, 9 out of every 10 people would have to be excluded, on average [1]. The guidelines most commonly followed to develop adult reference intervals, CLSI EP28-A3c, are essentially silent on how to address concerns specific to elderly populations. Analogous to the developmental processes that occur in the pediatric population, the geriatric population undergoes the physiological process of aging, which has a large impact on circulating biochemical markers in the elderly [96]. Geriatric patients often have one or more morbidities, are prescribed more than one medication, and vary widely in their physical ability, which all affect their biological state [97]. These factors require extensive partitioning to account for the differences in analyte levels, necessitating an even larger minimum sample size.

In an elderly population, it is difficult to decipher the extent of the apparent increase in biological variation that is part of the "normal" aging process [98]. For example, creatinine concentrations increase with age [12,99], but limits have not been established for what would be considered a normal increase rather than an increase that is indicative of disease. This represents a challenge in the process of outlier removal, since these outliers may represent natural variability within a specific group of individuals. This variability could be assessed with a statistical method, such as a sensitivity analysis to determine how influential outlying observations are on reference interval estimates [100]. As previously mentioned, geriatric subjects commonly have one or more comorbidities or are taking multiple medications, and therefore data may represent a patient population, rather than a healthy community representative sample. Although it is not recommended by CLSI to use patient data, it may be a useful estimate for this complex population. If hospital data sets are used, an indirect method (e.g., Hoffmann, Bhattacharyya, or Arzideh method) should be used to select apparently "healthy" individuals from the data set to ensure reference intervals are reliable and valid.

Typically, adult reference intervals are partitioned according to age groups, which show homogeneity in the analyte of interest. However, chronological age partitioning may not be suitable for elderly populations as one's chronological age is often not indicative of their biological state. Therefore using the characteristic approach to age partitioning could lead to grouping heterogeneous health states together. A more useful method would be to group individuals that have similar biological states through visually assessing the data to identify groups with similar analyte levels. This could be done by creating scatter or box plots against age, sex, number of morbidity, and/or type of morbidity [96]. The same statistical methods mentioned above in "Partitioning and minimum sample size" can be used here to test significant differences between the partitions. An additional challenge faced when partitioning geriatric reference intervals is due to difficulties in obtaining large enough sample sizes for these age groups. Partitions can be appropriately set based on a statistical test, but the sample sizes in the partitions may be inadequate. On the other hand, two very different groups may be inappropriately combined with the single purpose of attaining the recommended sample sizes. In both of these cases, the differences in the partitioning factor (i.e., age, biological state, and so on) are not adequately portrayed.

One ongoing initiative for geriatric reference intervals is the Swiss Senior Labor study (www.seniorlabor.ch),

which involves nearly 1500 elderly participants (60 years or older) to develop geriatric-specific reference intervals for various analytes including vitamin D [25(OH)], vitamin B12, homocysteine, and folate [101,102]. This group used CLSI EP28-A3c guidelines to establish reference intervals, specifically the nonparametric method as the sample size was large and the data were skewed. Exclusion criteria used, specific for classifying elderly patients as "healthy," excluded patients with clinical history suggestive of cognitive impairment as well as survival of less than 1 year following the study [101]. A study in 2012 reported reference values for 31 routine laboratory tests in a population of subjects aged 75 years [103]. Recognizing the challenges of using very strict criteria for defining a "healthy" population and the tradeoff with sufficient sample size, this study excluded all individuals with diabetes mellitus and then calculated reference values for the whole group and also for subgroups without cardiovascular disease [103].

Another study was conducted in 2013 to establish reference intervals for serum creatinine in the healthy geriatric population, using 3640 healthy Chinese individuals aged 60−89 years [104]. This study excluded participants on the basis of several factors including a diagnosed acute or chronic disease, undergone surgery in past 6 months, use of prescription medication within the previous 2 weeks, high blood pressure for over 3 years, and structural abnormalities of heart, liver, lungs, or kidney tested by ultrasonic wave, electrocardiogram, or chest X-ray film [104]. A recent Canadian initiative used data from the CHMS (a survey of Statistics Canada) to establish reference intervals for Canadians aged 3−80 years for common biochemical assays, immunoassays, and hematology assays [12−14]. The same exclusion criteria (including use of prescription medication or diagnosed serious medical or chronic illness) was used for all subjects, regardless of age, which resulted in approximately 70% of subjects aged 40−79 years excluded. A meta-analysis of 64 studies on geriatric reference intervals [96] showed extensive gaps in the reporting of statistical methods used for reference interval calculation. This observation highlights the need to apply rigorously current guidelines to reference interval determination for geriatric populations. Another promising new initiative is the Canadian Longitudinal Study on Aging (CLSA), which is a large, national, long-term study that will follow approximately 50,000 men and women between the ages of 45 and 85 for at least 20 years (www.clsa-elcv.ca/about-study). CLSA plans to establish a comprehensive geriatric reference interval database.

Reference interval transference and verification

When reference intervals are established, they are specific for the instrument, method, and population from which they were determined. To ensure all laboratories have access to reference intervals that are reliable for their specific procedures and populations without performing a full reference interval study, transference methods may be used to adapt previously established reference intervals.

Methodological approach to transference

For individual laboratories, determining reliable reference intervals for every new analytical test and method is often beyond their capabilities, making it very convenient and cost-effective if reference intervals established in another laboratory could be used. For reference intervals established in one laboratory (donor) to be used in another laboratory (receiver), the fulfillment of certain conditions is required. These include (1) the reference interval to be transferred must have been obtained properly and its generation must be fully documented; (2) the analytical systems must be comparable; and (3) the test subject populations must be comparable [7,105]. Transferring reference intervals, as opposed to performing an entirely new study, poses a major advantage of obviating the need for the laboratory to obtain samples from a large number of reference individuals.

The method of reference interval transference is the calculation of a new reference interval for a different laboratory based on an existing reference interval using the regression equation from a method comparison. First, fresh samples from a patient population are used to investigate the relationship between the analytical methods. The correlation between the analytical systems needs to be sufficiently high ($r^2 > 0.70$) in order to use regression statistics to calculate a new reference interval from the existing one [105]. Even if the analytical systems correlate extremely well ($r^2 > 0.9$), there may still be a significant difference between the laboratory results measured on the two systems due to bias [105]. For regression methods to be used sufficiently, laboratory test values must have a large enough range ratio and the intercept should be small in relation to the reference interval [105]. If the methods correlate sufficiently, linear regression is used to calculate the best fit regression line. Finally, the best fit regression line is used to calculate the transferred reference interval based on the existing reference interval. An example of an acceptable method comparison to allow reference interval transference is shown in Fig. 3.3. In this example, the analytical systems correlate extremely well ($r = 0.988$), the slope (1.02) is close to 1, and the y-intercept (−0.15) is close to 0. Furthermore, the reference interval to be transferred was 4.2−10.2, and the y-intercept (−0.15) would be sufficiently small. The final recommendation for a laboratory to adopt potentially transferred reference intervals is for the populations of both laboratories to be homogenous, meaning there are no major ethnic, social, or environmental differences among them [7]. Complete demographic information on the

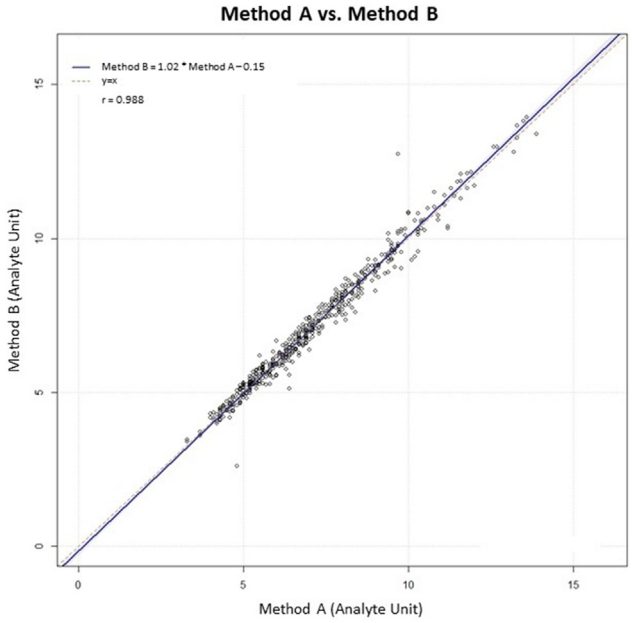


FIGURE 3.3 Graphical representation of an acceptable method comparison for transferring reference intervals. Values for a given analyte obtained by Method A (used to derive reference interval) strongly correlates ($r = 0.988$) with those obtained by Method B. A linear regression line acceptable for transferring the reference interval is obtained, demonstrated by minimal scatter around the regression line, a slope close to 1.0, and y-intercept close to 0.

original reference sample groups must be available and must correspond to the demographics of the new population [105]. However, it is not always practical for populations of both laboratories to be homogenous. The next step in adopting reference intervals from another laboratory is to verify the transferred reference intervals. If the transferred reference intervals are successfully verified using the receiving laboratories' local population, the populations from both laboratories can be assumed to be homogenous.

Limitations of the transference method

There are some caveats of using the transference method to establish reference intervals for a laboratory, rather than performing a new reference interval study. First, it is critically important to use an appropriate distribution of values. If a sufficient range is not represented, the correlation may appear worse than it is. Alternatively, if the range of samples is too large, the quality of the correlation can be overestimated. Second, the magnitude of the intercept calculated through linear regression must be interpreted based on how it compares to the range of data and the reference interval. If the intercept is large compared to the reference interval, the method bias may negate the appropriateness of transferring. Particularly, for immunoassays that are not traceable, if the confidence interval for the y-intercept does not encompass zero, it can be difficult to remove bias across the entire measuring range by simply using a slope correction. Finally, for particular analytes, linear regression may not be the most appropriate method to compare the two sets of data. Correction of the mean bias between the methods may be a better approach for analytes with a narrow range of values or discrete integers [7].

Transference initiatives

In 2013 a group from Yamaguchi University Graduate School of Medicine in Japan published results of a transference study for nonstandardized analytes in an Asian population [106]. First, common reference intervals for nonstandardized analytes in an Asian population were determined using centralized measurements [106], meaning that all specimens were transported to central laboratories, and therefore a single reagent and platform were used to eliminate assay-related variations. To transfer these reference intervals to other laboratories, specimens were measured on the central laboratory system as well as each of 37 local laboratory systems. Reference intervals were transferred using a calculated regression line from method comparison analysis. The coefficient of variation of slope b [CV(b)] was used to express the error in converting reference intervals using the regression line. A CV(b) of 10% was set as the cutoff value to allow the reference interval to be transferred. A limitation to this study, however, was that <40 samples were used for comparison in most cases, which is less than recommended in the CLSI guideline [7].

The CALIPER initiative has also performed a series of transference studies to broaden the applicability of their pediatric reference interval database [21,22,26,27,30]. CALIPER reference intervals were initially established on a single analytical system, the Abbott ARCHITECT assay platform. Transference studies were completed to address this limitation and allow these reference intervals to be applicable to four other major analytical systems, Beckman Coulter [22,26,27], Ortho [22,30], Roche [21,22], and Siemens [22]. Briefly, to perform transference, a method comparison was performed by analyzing patient samples on the original assay platform used to establish reference intervals and the platform for which transferred reference intervals were being calculated. If the methods correlated strongly ($r^2 > 0.70$), regression was used to determine the line of best fit. However, if methods were poorly correlated ($r^2 < 0.70$), the reference intervals could not be reliably transferred. The appropriateness of the linear model was assessed using quantile−quantile (Q−Q), studentized residual, and Bland−Altman plots. If the criteria of these three graphs were met, the line of best fit was used to calculate the transferred reference interval. The calculated 95% confidence intervals around each limit were used as limits with

which to verify the transferred reference intervals. Verification was then performed by analyzing 100 local reference samples that adequately encompassed as many age and sex partitions as possible. Reference intervals were considered verified when >90% of the reference samples fell within the transferred reference intervals, inclusive of the 95% confidence intervals. Although these studies provide reference intervals transferred to additional analytical platforms, these transference data are for specific assays, and the data do not validate reference intervals for individual analyzers, specific populations, or geographic locations.

Transference studies are invaluable to laboratory medicine to ensure robustly established reference intervals can be used by laboratories nationally and internationally, regardless of the analytical instrument they use. However, before implementing a reference interval into clinical practice, the newly transferred reference interval or a preexisting reference interval must be validated to ensure the donor and receiver laboratory populations are the same and there are no instrument biases.

Verification of transferred reference intervals

Three approaches can be used to assess if a preexisting or transferred reference interval can be accepted for use in a receiving laboratory: (1) a subjective assessment; (2) a statistical test on a relatively small number of reference individuals; and (3) an evaluation of a larger number of reference individuals.

A subjective assessment of the acceptability of the transferred reference intervals can be performed by ensuring all pertinent factors of the original reference interval study are consistent with the receiving laboratory's practice and test subject population [7]. This acceptability is based on an expert opinion following a careful examination of the conditions by which the reference interval was initially determined [105]. The pertinent factors must include the demographic variables and geographic location of the reference population, the preanalytical and analytical procedures, analytical performance, and the method of estimating reference intervals [7]. All factors must correspond to those in the receiving laboratory. This is an extremely subjective procedure, which poses many risks, and thus the second and third approaches are more commonly used.

The second approach is to assess a small number of reference individuals ($n = 20$) from the receiving laboratory's local healthy population. When using this approach, however, the analytical and preanalytical factors of the original reference value study still need to be consistent with the receiving laboratory's procedures. First, outliers need to be eliminated using either the Reed/Dixon [49,107] or Tukey [50] methods and new local reference samples must be obtained to replace them. If no more than two of the 20 reference subjects' values (or 10% of the test results) lie outside the preexisting or transferred 95% reference limits, this reference interval is considered valid for application in the receiving laboratory. If three or four test results exceed these limits, an additional 20 local reference samples must be obtained, with outliers once again removed. If no more than two of these new test results lie outside the reference limits, this reference interval is acceptable for use in the receiving laboratory. On the other hand, if three or more exceed the limits again (or if five or more in the original sample of 20 fall outside the limits), the analytical procedures used and biological characteristics of the populations must be reexamined and/or the receiving laboratory should establish its own reference intervals through a full-scale study [7]. This is a robust validation procedure, with the probability of false rejection of a reference interval <1% when one or more sets of 20 local reference samples are used [7]. However, if the preexisting or transferred reference interval is too wide for the receiving laboratory's local population, a false-negative result may occur. This highlights the criticality of ensuring the methods and populations are comparable between the laboratories prior to performing validation.

The third approach that can be used is similar to the second approach, but more extensive. This approach is most often used for analytes with reference intervals that are critically important for local clinical interpretation of the assay. In this approach, a larger population of reference individuals (e.g., $n = 60$) from the receiving laboratory's local population is assessed. Once again, the analytical and preanalytical factors of the original reference interval study need to be consistent with the receiving laboratory's procedures. Performing this larger study will have more statistical power for discovering differences between the original donor laboratory and the receiving laboratory's subject population.

The methods of transference and verification greatly improve the applicability of reference intervals to medical laboratories using different analytical instruments. A new focus is now harmonization in laboratory medicine, including harmonization of reference intervals, to standardize test result interpretation and increase transferability of results between the laboratories.

Toward reference interval harmonization

Harmonization in laboratory medicine is achieved through addressing all aspects of the total testing process (TTP) to ensure comparability and concordance of data obtained by different medical laboratories. The scope of harmonization in laboratory medicine is widespread to include aspects of the preanalytical, analytical, as well as the postanalytical phases, including test requests, sample collection, sample handling and transportation, tests and test profiles, terminology and units, reporting formats,

reference intervals and decision limits, and criteria for interpretation. Harmonization is fundamental to the quality of laboratory medicine with the ultimate goal of improving the accuracy and consistency of laboratory medicine and benefit patient health care and outcomes [108]. Many patients and physicians assume that they will receive the same, or at least comparable, result from a laboratory test measured by different laboratories at different times on the same sample and that results are interpreted accurately and consistently [109]. However, this is unfortunately not often the case, as many laboratory test results remain highly variable and poorly harmonized. Those requesting laboratory tests, receiving laboratory reports, developing information systems, and even laboratorians and physicians may be unaware of these critical differences, especially when results are transferred between the laboratories and differences in reporting units or methodology are not clarified. If clinicians assume these results can be directly compared, there is great potential for misinterpretation of laboratory test results and, consequently, erroneous diagnosis and/or treatment. The main drivers for laboratory medicine harmonization are first and foremost patient safety and the need for requesting physicians to be able to compare directly results from different clinical laboratories. As we are moving toward full electronic reporting of laboratory results to enable them to be viewed by a range of users, there is a greater need for harmonizing test names, units, and reference intervals associated with results.

Numerous studies have shown that the reference intervals for clinical analytes have larger variation than the analytical inaccuracy of their measurement [110,111]. Most recently, a national survey created and disseminated by the Canadian Society of Clinical Chemists (CSCC) hRI Working Group reported that reference interval variation was greater than test result variation for the majority of analytes assessed [112]. Furthermore, reference interval variation was particularly prominent in the pediatric population and varied substantially even between laboratories using the same analytical instrumentation. Therefore if the same patient sample is measured in two different laboratories using the same assay but different reference intervals, the patient's laboratory test result will be interpreted differently and may result in inappropriate diagnosis and/or treatment. In some cases, there are sound scientific reasons for differences in reference intervals, such as ethnicity and analytical methodology; however, both the UK Pathology Harmony study and CSCC hRI Working Group state that laboratories using the same instrument and reagents were often using different reference intervals without a sound explanation [112,113]. One way to mitigate this problem is to use the same reference interval or common reference interval. hRI is the use of a common reference interval across different analytical platforms and/or assays for a specified analyte. The ideal conditions to use a harmonized reference interval are where sound calibration and traceability are in place and where bias would not prevent the use of a common reference interval as evidenced from a between-method comparison. Assays that are traceable to reference measurement systems will be most capable of achieving harmonization, as it is the analytical quality of the assays that determine which analytes have the ability to share harmonized reference intervals (RIs). Several countries and regions have developed common RIs through either direct studies or using a consensus process. An organized plan and structured procedure are important to derive and validate common reference intervals, which will eventually be used to achieve a major national change in pathology reference intervals. Procedures developed and used by a few national initiatives are described in the next section.

One of the earliest hRI initiatives is the NORIP, which has established common reference intervals for 25 clinical chemistry analytes using a reference population of apparently healthy adults from five Nordic countries [33]. In this study, 102 participating laboratories collected blood samples from at least 25 healthy reference individuals and analyzed 25 commonly requested serum/plasma components from each reference individual. A reference material (control; consisting of fresh frozen liquid serum with values traceable to reference methods) was analyzed together with other serum pool controls in the same series as the project samples. Analytical data, method data, and data describing the reference individuals were submitted to a central database for evaluation and calculation of common reference intervals for the Nordic countries. The Pathology Harmony approach in England had an initial focus on clinical biochemistry reference intervals with a subsequent expansion to hematology and immunology [113]. Their harmonization technique used a three-phase process. First, they identified the variation and looked for scientific reasons supporting these differences. If a valid scientific basis was found for variation, the analyte was not harmonized. If there was no underlying evidence to explain identified variation, the second step was to review all reference interval data for the specific analyte and reach a consensus through expert group discussions. Finally, a harmonized reference interval was established and implemented into laboratories.

There have also been efforts by the Australasian Association of Clinical Biochemists (AACB) to encourage and assist laboratories to achieve hRI for common chemistry analytes where sound calibration and traceability are in place. The AACB and the Royal College of Pathologists of Australasia have recommended the implementation of a first panel of common reference intervals for use in Australia (Australasian Harmonised Reference Intervals for Adults and Australasian Harmonised Reference Intervals for Paediatrics)

[32]. The Australasian approach consisted of a checklist assessment process to assess the evidence for the use of common reference intervals, based on the following criteria: (1) define the analyte; (2) define the assays used, accuracy base, analytical specificity, and method-based bias; (3) consider important preanalytical differences and actions in response to interference; (4) define the principle behind the reference interval (e.g., central 95%); (5) describe the evidence for selection of common reference intervals [data sources (literature, lab surveys, and manufacturers), data mining, and allowable bias goal as the quality criterion for acceptance]; (6) consider partitioning based on age, sex, etc.; (7) define degree of rounding; (8) assess clinical considerations of the reference interval and flagging rates; (9) consider use of common reference interval; and (10) document and implement [32]. The Japanese Association of Medical Technologists (JAMT) has also established common reference intervals for 27 serum analytes for which certified reference materials are available and nine analytes frequently measured in routine tests. A hundred laboratories collaborated to recruit, sample, and measure the analytes of 3371 subjects aged 18−65 years. Allowable limits of trueness and intermediate precision based on JAMT criteria were applied to all reference values measured in these laboratories, and those within the allowable limits were used to establish common reference intervals. A characteristic feature of this study is the establishment of reference intervals by the collaboration of many core laboratories certified for metrological traceability nationwide [114].

The extensive effort to derive harmonized reference values for consistent and reliable interpretation of laboratory test results is wasted without effective communication of the information to end users. Laboratory acceptance should be sought at a national level prior to introducing common reference intervals. To assess the likely adoption rates for common reference intervals, a survey can be used to determine whether the laboratory is using a common reference interval already, would accept the common reference interval, or collect information as to their reason why they would not accept the common reference interval. National acceptance of new pathology reference intervals requires ongoing discussions by all involved stakeholders. In addition, criteria must be in place for each laboratory planning to adopt common reference intervals and use them in their clinical practice. Clinical laboratories can implement common reference intervals given there is verification of similar preanalytical conditions, traceability of the analytical method used, and similar characteristics of the tested population [115]. Although establishing and implementing harmonized reference intervals are critically important and will be of extreme value, continuing work is required. Consultation with clinical societies and education of local clinicians is critical if the new common reference intervals are to be used. Maximum benefit from harmonized reference limits will only be realized when all aspects of the TTP are harmonized.

Limitations of population-based reference intervals

Although population-based reference intervals are widely accepted and have numerous advantages for routine clinical application, there are some limitations that should be noted. First, it is important to consider the population from which the reference interval was derived. Differences between the reference population and the individual being tested (e.g., in terms of ethnicity) have the potential to cause misleading clinical interpretation of the test results. As 95% reference limits are normally calculated, laboratories should also be aware that 5% of healthy subjects will remain outside the reference interval and may subsequently be classified as "abnormal." Similarly, it is possible for individuals with subclinical or undetectable disease to be included in the calculation of reference intervals; therefore reference intervals may not necessarily be representative of a truly healthy population. Finally, the usefulness of a population-based reference interval may be evaluated by assessing the index of individuality (II), determined by the ratio of within-subject (CV_I) to between-subject (CV_G) biological variation ($II = CV_I/CV_G$). Analytes with a low II (≤ 1) have significant variation in analyte concentrations between individuals, and thus population-based reference intervals for these analytes may be less sensitive. For analytes with a high II (≥ 1), there is more variation within subjects than between subjects, and population-based reference intervals are more sensitive and, thus, more appropriate for these analytes.

Concluding remarks

In conclusion, reference intervals are central to laboratory medicine, serving as "health-associated benchmarks" by which clinical laboratory test results are compared and interpreted as either normal or abnormal. Reliable and accurate reference intervals have been severely lacking, especially in the pediatric and geriatric populations, due to the complexity of their establishment and the need to continuously update databases, as new assays and technologies are developed. Several aspects need to be considered when establishing reference intervals including selecting appropriate reference individuals and using the appropriate method and questionnaire to do so, preanalytical and analytical factors when handling and analyzing samples, exclusion criteria and outlier removal, partitioning of reference intervals based on key covariates, and selection of the appropriate statistical method to compute reference intervals.

As this process is time-consuming, costly, and often beyond the capabilities of individual laboratories, reference

intervals can be transferred from one laboratory (from which the reference interval was established) to another laboratory. Transferring reference intervals involves the calculation of a new reference interval for a different laboratory or analytical instrument by using the regression equation from a method comparison. Following transference, verification is used to assess if the transferred reference interval or a preexisting reference interval can be accepted for use in a receiving laboratory. Ensuring comparability and concordance of data obtained by different medical laboratories has been a recent central focus in laboratory medicine and can be achieved through harmonization. Harmonization of the TTP, including harmonization of reference intervals, will standardize test result interpretation, increase transferability of results between laboratories, and ultimately lead to successful implementation of harmonized reference intervals into clinical practice.

References

[1] P.S. Horn, A.J. Pesce, Reference intervals: an update, Clin. Chim. Acta Int. J. Clin. Chem. 334 (1–2) (2003) 5–23.

[2] B. Jung, K. Adeli, Clinical laboratory reference intervals in pediatrics: the CALIPER initiative, Clin. Biochem. 42 (16–17) (2009) 1589–1595.

[3] T. Badrick, Evidence-based laboratory medicine, Clin. Biochem. Rev. 34 (2) (2013) 43–46.

[4] M.J. Hallworth, P.L. Epner, C. Ebert, et al., Current evidence and future perspectives on the effective practice of patient-centered laboratory medicine, Clin. Chem. 61 (4) (2015) 589–599.

[5] J.L.V. Shaw, T. Binesh Marvasti, D. Colantonio, K. Adeli, Pediatric reference intervals: challenges and recent initiatives, Crit. Rev. Clin. Lab. Sci. 50 (2) (2013) 37–50.

[6] G. Jones, A. Barker, Reference intervals, Clin. Biochem. Rev. 29 (Suppl. 1) (2008) S93–S97.

[7] CLSI. Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guidelines - Third Edition. CLSI document EP28-A3c. Wayne, PA: Clinical and Laboratory Standards Institute; 2010.

[8] J.C. Boyd, Defining laboratory reference values and decision limits: populations, intervals, and interpretations, Asian J. Androl. 12 (1) (2010) 83–90.

[9] K. Thygesen, J. Mair, E. Giannitsis, et al., How to use high-sensitivity cardiac troponins in acute cardiac care, Eur. Heart J. 33 (18) (2012) 2252–2257.

[10] ISO 15189. Medical laboratories - requirements for quality and competence 2012. [Internet]. ISO. Available from: <http://www.iso.org/iso/catalogue_detail?csnumber = 56115> (cited 09. 07.18).

[11] M. Aytekin, K. Emerk, Accurate reference intervals are required for accurate diagnosis and monitoring of patients, EJIFCC 19 (2) (2008) 137–141.

[12] K. Adeli, V. Higgins, M. Nieuwesteeg, et al., Biochemical marker reference values across pediatric, adult, and geriatric ages: establishment of robust pediatric and adult reference intervals on the basis of the Canadian Health Measures Survey, Clin. Chem. 61 (8) (2015) 1049–1062.

[13] K. Adeli, V. Higgins, M. Nieuwesteeg, et al., Complex reference values for endocrine and special chemistry biomarkers across pediatric, adult, and geriatric ages: establishment of robust pediatric and adult reference intervals on the basis of the Canadian Health Measures Survey, Clin. Chem. 61 (8) (2015) 1063–1074.

[14] K. Adeli, J.E. Raizman, Y. Chen, et al., Complex biological profile of hematologic markers across pediatric, adult, and geriatric ages: establishment of robust pediatric and adult reference intervals on the basis of the Canadian Health Measures Survey, Clin. Chem. 61 (8) (2015) 1075–1086.

[15] T. Teodoro-Morrison, L. Kyriakopoulou, Y.K. Chen, et al., Dynamic biological changes in metabolic disease biomarkers in childhood and adolescence: a CALIPER study of healthy community children, Clin. Biochem. 48 (13–14) (2015) 828–836.

[16] J.E. Raizman, A.H. Cohen, T. Teodoro-Morrison, et al., Pediatric reference value distributions for vitamins A and E in the CALIPER cohort and establishment of age-stratified reference intervals, Clin. Biochem. 47 (9) (2014) 812–815.

[17] J.E. Raizman, F. Quinn, D.A. Armbruster, K. Adeli, Pediatric reference intervals for calculated free testosterone, bioavailable testosterone and free androgen index in the CALIPER cohort, Clin. Chem. Lab. Med. 53 (10) (2015) e239–e243.

[18] D. Konforte, J.L. Shea, L. Kyriakopoulou, et al., Complex biological pattern of fertility hormones in children and adolescents: a study of healthy children from the CALIPER cohort and establishment of pediatric reference intervals, Clin. Chem. 59 (8) (2013) 1215–1227.

[19] J. Kelly, J.E. Raizman, V. Bevilacqua, et al., Complex reference value distributions and partitioned reference intervals across the pediatric age range for 14 specialized biochemical markers in the CALIPER cohort of healthy community children and adolescents, Clin. Chim. Acta Int. J. Clin. Chem. 450 (2015) 196–202.

[20] K. Karbasy, D.C.C. Lin, A. Stoianov, et al., Pediatric reference value distributions and covariate-stratified reference intervals for 29 endocrine and special chemistry biomarkers on the Beckman Coulter Immunoassay Systems: a CALIPER study of healthy community children, Clin. Chem. Lab. Med. 54 (4) (2016) 643–657.

[21] V. Higgins, M.K. Chan, M. Nieuwesteeg, et al., Transference of CALIPER pediatric reference intervals to biochemical assays on the Roche Cobas 6000 and the Roche Modular P, Clin. Biochem. 49 (1–2) (2016) 139–149.

[22] M.P. Estey, A.H. Cohen, D.A. Colantonio, et al., CLSI-based transference of the CALIPER database of pediatric reference intervals from Abbott to Beckman, Ortho, Roche and Siemens Clinical Chemistry Assays: direct validation using reference samples from the CALIPER cohort, Clin. Biochem. 46 (13–14) (2013) 1197–1219.

[23] D.A. Colantonio, L. Kyriakopoulou, M.K. Chan, et al., Closing the gaps in pediatric laboratory reference intervals: a CALIPER database of 40 biochemical markers in a healthy and multiethnic population of children, Clin. Chem. 58 (5) (2012) 854–868.

[24] V. Bevilacqua, M.K. Chan, Y. Chen, D. Armbruster, B. Schodin, K. Adeli, Pediatric population reference value distributions for cancer biomarkers and covariate-stratified reference intervals in the CALIPER cohort, Clin. Chem. 60 (12) (2014) 1532–1542.

[25] D. Bailey, D. Colantonio, L. Kyriakopoulou, et al., Marked biological variance in endocrine and biochemical markers in childhood: establishment of pediatric reference intervals using healthy

community children from the CALIPER cohort, Clin. Chem. 59 (9) (2013) 1393–1405.
[26] P.A.T. Araújo, D. Thomas, T. Sadeghieh, et al., CLSI-based transference of the CALIPER database of pediatric reference intervals to Beckman Coulter DxC biochemical assays, Clin. Biochem. 48 (13–14) (2015) 870–880.
[27] M. Abou El Hassan, A. Stoianov, P.A.T. Araújo, et al., CLSI-based transference of CALIPER pediatric reference intervals to Beckman Coulter AU biochemical assays, Clin. Biochem. 48 (16–17) (2015) 1151–1159.
[28] V. Higgins, M.K. Chan, K. Adeli, Pediatric reference intervals for transferrin saturation in the CALIPER cohort, EJIFCC 28 (1) (2017) 77–84.
[29] V. Higgins, A.W.S. Fung, M.K. Chan, J. Macri, K. Adeli, Pediatric reference intervals for 29 Ortho VITROS 5600 immunoassays using the CALIPER cohort of healthy children and adolescents, Clin. Chem. Lab. Med. (2017).
[30] V. Higgins, D. Truong, A. Woroch, M.K. Chan, H. Tahmasebi, K. Adeli, CLSI-based transference and verification of CALIPER pediatric reference intervals for 29 Ortho VITROS 5600 chemistry assays, Clin. Biochem. 53 (2018) 93–103.
[31] V. Higgins, D. Truong, N.M.A. White-Al Habeeb, A.W.S. Fung, B. Hoffman, K. Adeli, Pediatric reference intervals for 1,25-dihydroxyvitamin D using the DiaSorin LIAISON XL assay in the healthy CALIPER cohort, Clin. Chem. Lab. Med. 56 (6) (2018) 964–972.
[32] J.R. Tate, K.A. Sikaris, G.R. Jones, et al., Harmonising adult and paediatric reference intervals in Australia and New Zealand: an evidence-based approach for establishing a first panel of chemistry analytes, Clin. Biochem. Rev. 35 (4) (2014) 213–235.
[33] P. Rustad, P. Felding, L. Franzson, et al., The Nordic Reference Interval Project 2000: recommended reference intervals for 25 common biochemical properties, Scand. J. Clin. Lab. Invest. 64 (4) (2004) 271–284.
[34] E. Grossi, R. Colombo, S. Cavuto, C. Franzini, The REALAB project: a new method for the formulation of reference intervals based on current data, Clin. Chem. 51 (7) (2005) 1232–1240.
[35] H.E. Solberg, International Federation of Clinical Chemistry (IFCC), Scientific Committee, Clinical Section, Expert Panel on Theory of Reference Values, and International Committee for Standardization in Haematology (ICSH), Standing Committee on Reference Values. Approved Recommendation (1986) on the theory of reference values. Part 1. The concept of reference values, J. Clin. Chem. Clin. Biochem. 25 (5) (1987) 337–342.
[36] C. PetitClerc, P. Wilding, International Federation of Clinical Chemistry (IFCC), Scientific Committee, Clinical Section. The theory of reference values. Part 2. Selection of individuals for the production of reference values, J. Clin. Chem. Clin. Biochem. 22 (2) (1984) 203–208.
[37] H.E. Solberg, C. PetitClerc, International Federation of Clinical Chemistry (IFCC), Scientific Committee, Clinical Section, Expert Panel on Theory of Reference Values. Approved recommendation (1988) on the theory of reference values. Part 3. Preparation of individuals and collection of specimens for the production of reference values, J. Clin. Chem. Clin. Biochem. 26 (9) (1988) 593–598.
[38] H.E. Solberg, D. Stamm, IFCC recommendation: the theory of reference values. Part 4. Control of analytical variation in the production, transfer and application of reference values, J. Autom. Chem. 13 (5) (1991) 231–234.
[39] H.E. Solberg, The theory of reference values Part 5. Statistical treatment of collected reference values. Determination of reference limits, J. Clin. Chem. Clin. Biochem. 21 (11) (1983) 749–760.
[40] R. Dybkaer, International federation of clinical chemistry (IFCC) 1),2 the theory of reference values. Part 6. Presentation of observed values related to reference values, J. Clin. Chem. Clin. Biochem. 20 (11) (1982) 841–845.
[41] H.E. Solberg, The IFCC recommendation on estimation of reference intervals. The RefVal program, Clin. Chem. Lab. Med. 42 (7) (2004) 710–714.
[42] K. Sikaris, Application of the Stockholm hierarchy to defining the quality of reference intervals and clinical decision limits, Clin. Biochem. Rev. 33 (4) (2012) 141–148.
[43] R.G. Hoffmann, Statistics in the practice of medicine, JAMA 185 (1963) 864–873.
[44] C.G. Bhattacharya, A simple method of resolution of a distribution into Gaussian components, Biometrics 23 (1) (1967) 115–135.
[45] J.B. Hemel, F.R. Hindriks, W. van der Slik, Critical discussion on a method for derivation of reference limits in clinical chemistry from a patient population, J. Autom. Chem. 7 (1) (1985) 20–30.
[46] F. Arzideh, W. Wosniok, E. Gurr, et al., A plea for intra-laboratory reference limits. Part 2. A bimodal retrospective concept for determining reference limits from intra-laboratory databases demonstrated by catalytic activity concentrations of enzymes, Clin. Chem. Lab. Med. 45 (8) (2007) 1043–1057.
[47] F. Arzideh, G. Brandhorst, E. Gurr, et al., An improved indirect approach for determining reference limits from intra-laboratory data bases exemplified by concentrations of electrolytes, Laboratoriums Medizin 33 (2) (2009) 52–66.
[48] D.B. Rorabacher, Statistical treatment for rejection of deviant values: critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidence level, Anal. Chem. 63 (2) (1991) 139–146.
[49] W. Dixon, Processing date for outliers, Biometrics 9 (1953) 74–89.
[50] J. Tukey, Exploratory Data Analysis, Addison-Wesley, Boston, MA, 1977.
[51] M. Hubert, Van, S. der Veeken, Outlier detection for skewed data, J. Chemom. 22 (3–4) (2008) 235–246.
[52] T.J. Sinton, D.M. Cowley, S.J. Bryant, Reference intervals for calcium, phosphate, and alkaline phosphatase as derived on the basis of multichannel-analyzer profiles, Clin. Chem. 32 (1 Pt 1) (1986).
[53] E.K. Harris, J.C. Boyd, On dividing reference data into subgroups to produce separate reference ranges, Clin. Chem. 36 (2) (1990) 265–270.
[54] A. Lahti, P.H. Petersen, J.C. Boyd, P. Rustad, P. Laake, H.E. Solberg, Partitioning of nonGaussian-distributed biochemical reference data into subgroups, Clin. Chem. 50 (5) (2004) 891–900.
[55] M.J. Healy, J. Rasbash, M. Yang, Distribution-free estimation of age-related centiles, Ann. Hum. Biol. 15 (1) (1988) 17–22.
[56] A. Virtanen, V. Kairisto, E. Uusipaikka, Regression-based reference limits: determination of sufficient sample size, Clin. Chem. 44 (11) (1998) 2353–2358.

[57] A. Virtanen, V. Kairisto, K. Irjala, A. Rajamäki, E. Uusipaikka, Regression-based reference limits and their reliability: example on hemoglobin during the first year of life, Clin. Chem. 44 (2) (1998) 327–335.
[58] F.E. Harrell, C.E. Davis, A new distribution-free quantile estimator, Biometrika 69 (3) (1982) 635–640.
[59] P. Horn, A. Pesce, Reference Intervals: A User's Guide, AACC Press, Washington, DC, 2005.
[60] E.K. Harris, D.L. DeMets, Estimation of normal ranges and cumulative proportions by transforming observed distributions to Gaussian form, Clin. Chem. 18 (7) (1972) 605–612.
[61] C.H. Daly, V. Higgins, K. Adeli, V.L. Grey, J.S. Hamid, Reference interval estimation: methodological comparison using extensive simulations and empirical data, Clin. Biochem. 50 (18) (2017) 1145–1158.
[62] R. Flatman, M. Legg, G.R. Jones, P. Graham, D. Moore, J. Tate, Recommendations for reporting and flagging of reference limits on pathology reports, Clin. Biochem. Rev. 35 (4) (2014) 199–202.
[63] P.S. Horn, A.J. Pesce, Effect of ethnicity on reference intervals, Clin. Chem. 48 (10) (2002) 1802–1804.
[64] E. Lim, J. Miyamura, J.J. Chen, Racial/ethnic-specific reference intervals for common laboratory tests: a comparison among Asians, Blacks, Hispanics, and White, Hawaii. J. Med. Public Health 74 (9) (2015) 302–310.
[65] H. Tahmasebi, K. Trajcevski, V. Higgins, K. Adeli, Influence of ethnicity on population reference values for biochemical markers, Crit. Rev. Clin. Lab. Sci. 55 (5) (2018) 359–375.
[66] B.H. Shirts, A.R. Wilson, B.R. Jackson, Partitioning reference intervals by use of genetic information, Clin. Chem. 57 (3) (2011) 475–481.
[67] M. Vincent-Viry, F. Schiele, R. Gueguen, K. Bohnet, S. Visvikis, G. Siest, Biological variations and genetic reference values for apolipoprotein E serum concentrations: results from the STANISLAS cohort study, Clin. Chem. 44 (5) (1998) 957–965.
[68] P. Shahabi, G. Siest, B. Herbeth, N.C. Ndiaye, S. Visvikis-Siest, Clinical necessity of partitioning of human plasma haptoglobin reference intervals by recently-discovered rs2000999, Clin. Chim. Acta. 413 (19–20) (2012) 1618–1624.
[69] Y. Ozarda, Reference intervals: current status, recent developments and future considerations, Biochem. Medica 26 (1) (2016) 5–16.
[70] M. de Onis, A.W. Onyango, E. Borghi, A. Siyam, C. Nishida, J. Siekmann, Development of a WHO growth reference for school-aged children and adolescents, Bull. World Health Organ. 85 (9) (2007) 660–667.
[71] R.J. Kuczmarski, C.L. Ogden, S.S. Guo, et al., 2000 CDC growth charts for the United States: methods and development, Vital. Health Stat. 11 (246)(2002) 1–190.
[72] K. England, C. Thorne, L. Pembrey, P.-A. Tovo, M.-L. Newell, Age- and sex-related reference ranges of alanine aminotransferase levels in children: European paediatric HCV network, J. Pediatr. Gastroenterol. Nutr. 49 (1) (2009) 71–77.
[73] K.P. Kohse, M. Thamm, KiGGS - the German survey on children's health as data base for reference intervals, Clin Biochem. 44 (2011) 479.
[74] C.R. McCudden, J. Brooks, P. Figurado, P.R. Bourque, Cerebrospinal fluid total protein reference intervals derived from 20 years of patient data, Clin. Chem. 63 (12) (2017) 1856–1865.
[75] J. Zierk, F. Arzideh, T. Rechenauer, et al., Age- and sex-specific dynamics in 22 hematologic and biochemical analytes from birth to adolescence, Clin. Chem. 61 (7) (2015) 964–973.
[76] J. Zierk, F. Arzideh, R. Haeckel, et al., Pediatric reference intervals for alkaline phosphatase, Clin. Chem. Lab. Med. 55 (1) (2017) 102–110.
[77] L. Mørkrid, A.D. Rowe, K.B.P. Elgstoen, et al., Continuous age- and sex-adjusted reference intervals of urinary markers for cerebral creatine deficiency syndromes: a novel approach to the definition of reference intervals, Clin. Chem. 61 (5) (2015) 760–768.
[78] J. Zierk, F. Arzideh, R. Haeckel, W. Rascher, M. Rauh, M. Metzler, Indirect determination of pediatric blood count reference intervals, Clin. Chem. Lab. Med. 51 (4) (2013) 863–872.
[79] F. Ceriotti, Establishing pediatric reference intervals: a challenging task, Clin. Chem. 58 (5) (2012) 808–810.
[80] P. Royston, D.G. Altman, Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling, Appl. Stat. 43 (3) (1994) 429.
[81] T.J. Cole, P.J. Green, Smoothing reference centile curves: the LMS method and penalized likelihood, Stat. Med. 11 (10) (1992) 1305–1319.
[82] A. Gannoun, S. Girard, C. Guinot, J. Saracco, Reference curves based on non-parametric quantile regression, Stat. Med. 21 (20) (2002) 3119–3135.
[83] Y. Wei, A. Pere, R. Koenker, X. He, Quantile regression methods for reference growth charts, Stat. Med. 25 (8) (2006) 1369–1382.
[84] K.P. Kohse, KiGGS - the German survey on children's health as data base for reference intervals and beyond, Clin. Biochem. 47 (9) (2014) 742–743.
[85] M. Lange, H.G. Butschalowsky, F. Jentsch, R. Kuhnert, A. Schaffrath Rosario, M. Schlaud, et al., [The first KiGGS follow-up (KiGGS Wave 1): study conduct, sample design, and response], Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 57 (7) (2014) 747–761.
[86] S.P. Wyness, W.L. Roberts, J.A. Straseski, Pediatric reference intervals for four serum bone markers using two automated immunoassays, Clin. Chim. Acta. 415 (2013) 169–172.
[87] M.M. Kushnir, A.L. Rockwood, W.L. Roberts, D. Abraham, A.N. Hoofnagle, A.W. Meikle, Measurement of thyroglobulin by liquid chromatography-tandem mass spectrometry in serum and plasma in the presence of antithyroglobulin autoantibodies, Clin. Chem. 59 (6) (2013) 982–990.
[88] M.M. Kushnir, T. Blamires, A.L. Rockwood, et al., Liquid chromatography-tandem mass spectrometry assay for androstenedione, dehydroepiandrosterone, and testosterone with pediatric and adult reference intervals, Clin. Chem. 56 (7) (2010) 1138–1147.
[89] M.M. Kushnir, A.L. Rockwood, J. Bergquist, et al., High-sensitivity tandem mass spectrometry assay for serum estrone and estradiol, Am. J. Clin. Pathol. 129 (4) (2008) 530–539.
[90] C.-N. Lin, A. Wilson, B.B. Church, S. Ehman, W.L. Roberts, G. A. McMillin, Pediatric reference intervals for serum copper and zinc, Clin. Chim. Acta. 413 (5–6) (2012) 612–615.
[91] S.M. Clifford, A.M. Bunker, J.R. Jacobsen, W.L. Roberts, Age and gender specific pediatric reference intervals for aldolase,

amylase, ceruloplasmin, creatine kinase, pancreatic amylase, prealbumin, and uric acid, Clin. Chim. Acta. 412 (9–10) (2011) 788–790.
[92] P.R. Slev, A.M. Bunker, W.E. Owen, W.L. Roberts, Pediatric reference intervals for random urine calcium, phosphorus and total protein, Pediatr. Nephrol. Berl. Ger. 25 (9) (2010) 1707–1710.
[93] A.W. Meikle, M.M. Kushnir, A.L. Rockwood, et al., Adrenal steroid concentrations in children seven to seventeen years of age, J. Pediatr. Endocrinol. Metab 20 (12) (2007) 1281–1291.
[94] L. Hilsted, P. Rustad, L. Aksglæde, K. Sørensen, A. Juul, Recommended Nordic paediatric reference intervals for 21 common biochemical properties, Scand. J. Clin. Lab. Invest. 73 (1) (2013) 1–9.
[95] S.I. Chaudhry, G. McAvay, Y. Ning, H.G. Allore, A.B. Newman, T. M. Gill, Geriatric impairments and disability: the cardiovascular health study, J. Am. Geriatr. Soc. 58 (9) (2010) 1686–1692.
[96] E. Arseneau, C.M. Balion, Statistical methods used in the calculation of geriatric reference intervals: a systematic review, Clin. Chem. Lab. Med. 54 (3) (2016) 377–388.
[97] W.R. Faulkner, S. Meites, Geriatric Clinical Chemistry: Reference Values, AACC Press, Washington, DC, 1994.
[98] A. Kruger, The limits of normality in elderly patients, Baillieres Clin. Haematol. 1 (2) (1987) 271–289.
[99] M. Andreassen, K. Nielsen, I. Raymond, L.Ø. Kristensen, J. Faber, Characteristics and reference ranges of insulin-like growth factor-I measured with a commercially available immunoassay in 724 healthy adult Caucasians, Scand. J. Clin. Lab. Invest. 69 (8) (2009) 880–885.
[100] C.H. Daly, X. Liu, V.L. Grey, J.S. Hamid, A systematic review of statistical methods used in constructing pediatric reference intervals, Clin. Biochem. 46 (13–14) (2013) 1220–1227.
[101] M. Risch, D.W. Meier, B. Sakem, et al., Vitamin B12 and folate levels in healthy Swiss senior citizens: a prospective study evaluating reference intervals and decision limits, BMC Geriatr. 15 (2015) 82.
[102] B. Sakem, C. Nock, Z. Stanga, et al., Serum concentrations of 25-hydroxyvitamin D and immunoglobulins in an older Swiss cohort: results of the Senior Labor Study, BMC Med. 11 (2013) 176.
[103] I. Ryden, L. Lind, A. Larsson, Reference values of thirty-one frequently used laboratory markers for 75-year-old males and females, Ups. J. Med. Sci. 117 (3) (2012) 264–272.
[104] M. Huang, J.-J. Yang, J.-J. Yang, A.-G. Tang, Reference intervals for serum creatinine levels in the healthy geriatric population, Clin. Biochem. 46 (15) (2013) 1419–1422.
[105] A. Geffré, K. Friedrichs, K. Harr, D. Concordet, C. Trumel, J.-P. Braun, Reference values: a review, Vet. Clin. Pathol. 38 (3) (2009) 288–298.
[106] K. Ichihara, F. Ceriotti, M. Kazuo, et al., The Asian project for collaborative derivation of reference intervals: (2) results of non-standardized analytes and transference of reference intervals to the participating laboratories on the basis of cross-comparison of test results, Clin. Chem. Lab. Med. 51 (7) (2013) 1443–1457.
[107] A.H. Reed, R.J. Henry, W.B. Mason, Influence of statistical method used on the resulting estimate of normal range, Clin. Chem. 17 (4) (1971) 275–284.
[108] A.K. Aarsand, S. Sandberg, How to achieve harmonisation of laboratory testing -the complete picture, Clin. Chim. Acta. 432 (2014) 8–14.
[109] M. Panteghini, Implementation of standardization in clinical practice: not always an easy task, Clin. Chem. Lab. Med. 50 (7) (2012) 1237–1241.
[110] G.R.D. Jones, A. Barker, J. Tate, C.-F. Lim, K. Robertson, The case for common reference intervals, Clin. Biochem. Rev. 25 (2) (2004) 99–104.
[111] G.R. Jones, Laboratory reporting of urine protein and albumin, Clin. Biochem. Rev. 32 (2) (2011) 103–107.
[112] K. Adeli, V. Higgins, D. Seccombe, et al., National survey of adult and pediatric reference intervals in clinical laboratories across Canada: a report of the CSCC working group on reference interval harmonization, Clin. Biochem. (2017).
[113] J. Berg, V. Lane, Pathology Harmony; a pragmatic and scientific approach to unfounded variation in the clinical laboratory, Ann. Clin. Biochem. 48 (3) (2011) 195–197.
[114] Y. Yamamoto, S. Hosogaya, S. Osawa, et al., Nationwide multicenter study aimed at the establishment of common reference intervals for standardized clinical laboratory tests in Japan, Clin. Chem. Lab. Med. 51 (8) (2013) 1663–1672.
[115] J.R. Tate, R. Johnson, J.H. Barth, M. Panteghini, "Harmonization of laboratory testing - A global activity", Clin. Chim. Acta. 432 (2014) 1–3.

Self-assessment questions

1. Which of the following statements are true regarding selection of reference individuals?
   a. According to CLSI, reference intervals should be based on indirect sampling of healthy individuals from the age group(s) of interest.
   b. Use of a posteriori selection criteria is not recommended by the CLSI.
   c. The quality of the reference interval is not affected by the reference population.
   d. Lack of appropriate inclusion/exclusion criteria can jeopardize the validity of the reference interval.
   e. In indirect sampling, partitioning takes place after sample collection and analyte testing.
2. Which of the following statement(s) are true regarding calculation of reference intervals?
   a. According to CLSI, a minimum of 120 reference individuals are needed per partition/subgroup.
   b. According to CLSI, only nonparametric rank order analysis should be used to calculate reference intervals.
   c. Reference intervals usually represent the central 98% of values from an apparently healthy reference population.
   d. Outliers should be excluded before reference intervals are calculated using either the Harris and Boyd or the Tukey method, as recommended by CLSI.
   e. The robust method by Horn and Pesce is recommended by the CLSI as the preferred method for calculating reference intervals by parametric analysis.
3. Which of the following is NOT an example of preanalytical variation?
   a. fasted versus nonfasted state of subjects
   b. tourniquet time
   c. separation of serum/plasma
   d. prior diet and stress level
   e. reagent lot-to-lot variability
4. Which of the following statements are NOT true regarding reference intervals and clinical decision limits?
   a. In some cases, only one side of the reference distribution is clinically significant.
   b. Reference limits represent cut points associated with disease.
   c. HbA1c is an example of an analyte for which clinical decision limits are used.
   d. For some analytes, the 99th percentile may be used to calculate reference intervals when greater specificity is required.
   e. For the majority of biochemical markers, reference intervals are used to interpret laboratory test results.
5. Which of the following is NOT true regarding partitioning of reference intervals?
   a. The Harris and Boyd method is one of the methods recommended by the CLSI.
   b. Partitions may be statistically valid but not clinically relevant.
   c. Age, sex, and ethnicity are examples of covariates that may necessitate reference interval partitioning.
   d. Construction of continuous reference intervals is never recommended by the CLSI.
   e. Partitions may be clinically relevant but not statistically valid.
6. Which of the following is NOT true regarding establishing reference intervals using the indirect method?
   a. The indirect method is applied to a population containing both healthy and pathological subjects.
   b. The majority of observations in the data set are assumed to be from healthy individuals.
   c. The traditional nonparametric method is most commonly used to establish indirect reference intervals.
   d. Indirect sampling may be necessary when recruitment of a large sample population is difficult.
   e. Indirect methods are used to define the data distribution from healthy subjects among pathological subjects.
7. Age-dependent reference curves:
   a. can be established using direct or indirect methods
   b. can more easily be implemented into laboratory information systems
   c. can be established using several statistical methods, including fractional polynomials and nonparametric quantile regression
   d. A and C
   e. all of the above
8. Which of the following is NOT true regarding pediatric reference intervals?
   a. The pediatric population differs from adults in regards to organ maturity, developmental stages, and immune responsiveness.
   b. The pediatric population always requires their own reference interval.
   c. Recruiting a sufficiently large and healthy pediatric reference population can be a challenge.
   d. Growth and sexual maturation during puberty can affect the levels of many analytes.

e. KiGGS, CALIPER, and CHILDx are regional initiatives developed to address gaps in pediatric reference intervals.

9. Which of the following is true regarding geriatric reference intervals?
- **a.** Geriatric reference values should be partitioned by chronological age, as this is indicative of their biological state.
- **b.** Unlike the pediatric population, it is not a challenge to recruit a sufficiently large and healthy geriatric reference population.
- **c.** There are specific CLSI guidelines with detailed recommendations to establish geriatric reference intervals.
- **d.** There are few studies on geriatric reference intervals and the statistical approaches used vary widely.
- **e.** Reference interval partitions are rarely needed in the geriatric population, as it is very homogenous.

10. Reference intervals are specific for the _____ from which they were determined:
- **a.** instrument
- **b.** analytical method
- **c.** population
- **d.** instrument and analytical method
- **e.** instrument, analytical method, and population

11. Which of the following is a condition that must be fulfilled for reference intervals established in one laboratory to be used in another laboratory?
- **a.** The reference interval to be transferred must have been obtained properly.
- **b.** The analytical systems must be comparable between laboratories.
- **c.** The test subject populations must be comparable.
- **d.** Full documents must be available on the generation of the reference interval to be transferred.
- **e.** All of the above conditions must be fulfilled.

12. If the r^2 value for the correlation between methods is less than _____, the reference intervals cannot be reliably transferred.
- **a.** 0.50
- **b.** 0.60
- **c.** 0.70
- **d.** 0.80
- **e.** 0.90

Answers

1. d
2. a
3. e
4. b
5. d
6. c
7. d
8. b
9. d
10. e
11. e
12. c

Chapter 4

Method validation

Athena K. Petrides[1,2], Jayme L. Dahlin[1,2], and Stacy E.F. Melanson[1,2]
[1]Department of Pathology, Brigham and Women's Hospital, Boston, MA, United States, [2]Harvard Medical School, Boston, MA, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Discuss the regulatory requirements associated with clinical method validation studies and the different organizations involved in generating these requirements.
- List the critical components of clinical method validation studies.
- Describe acceptance and rejection criteria for clinical method validation studies.
- Define a laboratory developed test and outline relevant validation parameters and processes.

Introduction

The evaluation of method performance characteristics is an essential responsibility of any clinical laboratory. Although now considered a rather heavily regulated routine function, it is one of the most important services laboratory medicine can offer to optimize patient care. This became widely evident and publicized in 1987 when the *New York Times* revealed that several undiagnosed cervical cancer deaths were linked to false-negative Pap smear results. In response to this report, the United States Congress voted for further oversight of any hospital, independent, or physician-owned clinical laboratory, which led to the Clinical Laboratory Improvement Amendments of 1988 (CLIA '88) being signed into US public law. Along with defining which analytical characteristics require assessment, this landmark federal law also states that laboratories must enroll in proficiency testing (see Chapter 15: Laboratory regulations and compliance), lists qualifications and functions for testing personnel, and enforces a quality control (QC) program.

Most assays performed in the clinical laboratory have already been approved by the US Food and Drug Administration (FDA), a process which requires the manufacturer to present data on their method validation. According to CLIA '88, clinical laboratories are only required to verify the manufacturer's claims by evaluating a limited set of criteria (Table 4.1). However, if the method is to be used differently than intended by the manufacturer, or if it has not been FDA-approved, it is then considered a modified test, and is held to the same regulatory standards as a laboratory developed test (LDT). In such cases, the clinical laboratory is responsible for evaluating the full set of performance criteria described in subsequent sections of this chapter. Even though federal law outlines the minimum requirements for method evaluation, state departments of public health may enforce stricter criteria.

Regulations

Even though law defines which performance characteristics should be evaluated, it neither includes acceptability criteria nor a standard procedure for assessing the listed characteristics. The National Committee on Clinical Laboratory Standards (NCCLS) was the first committee assembled to create standards in laboratory testing. Their first publication in 1969 on how to create manuals for laboratory instruments was endorsed by many agencies, including the FDA, the College of American Pathologists (CAP), and the Centers for Disease Control and Prevention. Another voluntary agency, the Clinical and Laboratory Standards Institute (CLSI), emerged in 1977 and focused primarily on standardizing reference materials and methods across laboratories. The two committees, NCCLS and CLSI, partnered in 2005 and are now known as CLSI. This agency has published over 300 standards that are solicited within the United States and internationally.

Accrediting agencies are now also issuing guidelines based on CLSI standards around key parameters for assessment as well as appropriate acceptability criteria. The Joint Commission, a nonprofit agency, offers an accreditation to both the laboratory and the hospital that is accepted nationwide as an indicator of quality. The CAP Laboratory Accreditation Program also accredits

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00004-1

TABLE 4.1 Regulatory requirements per organization/ guidelines.

Standard	CLIA	CAP	TJC	FDA
Waived testing				
Verification/validation studies are not required				
Moderately complex (verification)				
Precision	X	X	X	n/a
Accuracy	X	X	X	
AMR	X	X	X	
Method comparison		X	X	
Reference range		X		
Carryover		X		
Laboratory developed tests (validation)				
Precision	X	X	X	X
Accuracy	X	X	X	X
AMR	X	X	X	X
Method comparison		X	X	X
Reference range	X	X		X
Carryover		X		X
Sensitivity	X	X	X	X
Specificity	X	X	X	X
Interferences	X	X	X	X
Stability				X
Dilution				X
Recovery				X
Selectivity				X

AMR, Analytical measuring range; *CAP*, College of American Pathologists; *CLIA*, Clinical Laboratory Improvement Amendments; *FDA*, Food and Drug Administration; *TJC*, The Joint Commission.

laboratories, often using a more rigorous and customized checklist than other agencies for FDA-approved tests. Both agencies refer to CLSI standards as evidence of compliance with respect to method evaluation. A comparison of requirements for method verification between the various agencies and committees issuing guidance documents is summarized in Table 4.1.

According to CLIA '88, testing falls into the following FDA-defined categories based on assay complexity: waived, moderately complex, and highly complex. Waived testing can be performed under a CLIA certificate of waiver, thus negating the need for a rigorous method verification process. However, each of the testing categories, including waived testing, has a different set of criteria for method verification, which can differ depending on the ruling committee or agency (Table 4.1).

Analytical goals and method selection

The need for method selection arises frequently in the clinical laboratory. A laboratory may be insourcing a send-out test or implementing a new assay to the market to enhance patient care. The laboratory may also be upgrading or switching instrumentation and selecting an analytical platform with multiple methodologies. Furthermore, troubleshooting may prompt the review of other available methods. For example, the clinical and/or analytical performance of an assay may no longer be adequate, requiring the laboratory to investigate alternatives. Importantly, the manufacturer may issue a notification stating that the laboratory must transition to a new assay formulation. In this scenario, the laboratory can choose to adopt the new formulation if the validation is acceptable, or switch assays. The manufacturer may also discontinue an assay, thus prompting the laboratory to determine whether the clinical utility warrants selecting a different method. Regardless of the scenario, the laboratory must research available assays and their performance as well as available instrumentation prior to performing method validation. In clinical practice, this process is often more time-consuming than the validation itself. An overview of definitions used in method evaluation and verification is provided in Table 4.2.

Utilizing established analytical goals for method selection is the critical first step in this process. It is recommended that laboratories define acceptability criteria, including analytical goals for all tests in the laboratory that can be used to guide method selection. In addition, parameters such as sensitivity and specificity (see Chapter 2: Statistical methods in laboratory medicine) are important to define, as testing with acceptable analytical performance may not necessarily meet the standards for clinical care. Instrumentation, cost, turnaround time, manufacturer reliability and service, informatics requirements, and other parameters should also be considered [1,2]. This chapter focuses on analytical acceptability and method validation. In addition, laboratories should not assume that the new method will be superior to the predicate method.

Multiple resources are available to help laboratories define analytical goals. For regulated analytes, many laboratories will use CLIA guidelines [3]. For nonregulated analytes, such as lactate and B-type natriuretic peptides, guidelines published by CAP or other proficiency testing programs are helpful in defining analytical benchmarks. Laboratories may also choose to discuss analytical goals with their clinician colleagues. Clinical goals ensure that the testing can accurately diagnose, monitor, and/or prevent disease, and can shape acceptable analytical performance criteria. However, this approach may be more

TABLE 4.2 Definition of common terms.

Term	Definition
Accuracy	The ability to recover the correct amount of analyte present in a specimen
AMR	The range of numeric results a method can produce without pretreatment not otherwise part of the usual analytic process (e.g., dilution); also referred to as "range" (CLIA)
Between-run precision	Assesses the imprecision of the assay over the long term (e.g., 10–20 days); also referred to as "day-to-day precision"
Calibration	Known standards are used to create a working curve to ensure the method produces accurate results
Carryover	The unintended transfer of substances (e.g., analyte, reagent, and diluent) between samples or assay compartments
CRR	The range of values an assay can quantitatively report with sample pretreatment such as dilution or concentration
CV	Measure of reproducibility of an assay either within a single run or between runs; CV = (SD/mean × 100)
Confidence interval	An interval that will contain the true value a certain percentage of the time
Constant bias	Results are always greater (or lesser) by the same amount; represented by b in the straight line equation ($y = mx + b$)
Correlation coefficient (R)	Specifies the degree of correlation not the degree to which the methods match. Ranges from −1 to +1 with result of 0 = no correlation and results of −1 or +1 as a perfect correlation
Intercept	Value of the y-variable when the x-variable has a value of zero
Interference	The modulation of the assay readout by nonanalyte substances in a sample (exogenous and/or endogenous); artifactual increase or decrease in apparent concentration or intensity of an analyte due to the presence of a substance that reacts nonspecifically with either the detecting reagent or the signal itself (CLIA)
Least-squares regression	The method of statistically placing the location of the estimated line or curve among the data, so that the sum of the squares of the distances of each data point from the line in the perpendicular direction from the x-axis is minimized
LOB	The highest apparent analyte concentration expected in blank sample replicates containing no analyte
LOD	The lowest analyte concentration likely to be reliably distinguished from the LOB and at which detection is feasible
LOQ	The lowest concentration at which the analyte cannot only be reliably detected but at which some predefined goals for bias and imprecision are met; may meet or exceed the LOD
Linearity	The ability of an assay to provide results that are directly proportional to analyte quantity in a test sample
Matrix effect	The direct or indirect alteration or interference in response due to the presence of unintended analytes or other interfering substances in a sample
Medical decision point	Value(s) that are clinically important for patient management
Method comparison	The testing of the same sample by two independent laboratory assays for the same analyte for the purpose of characterizing test inaccuracy and systemic errors
Outlier	A result that exceeds 10 times the standard error of estimate
Parallelism	Relative accuracy from recovery tests for endogenous analytes in a biological matrix compared with calibrators
Precision	Ability to obtain the same result upon repeated measurement of a specimen
Proportional bias	Results are always higher or lower by the same proportion; represented by m in the equation of a straight line ($y = mx + b$)
Random error	An error caused by factors that vary from one measurement to the next measurement
Recovery	Hundred times measured mean/assigned value; utilized in reportable range experiments
Reportable range	The span of test result values over which the laboratory can establish or verify the accuracy of the instrument or test system measurement response

(*Continued*)

TABLE 4.2 (Continued)

Term	Definition
Selectivity	The ability of the bioanalytical method to measure and differentiate analytes in the presence of components that may be expected to be present (includes metabolites, impurities, degradants, or matrix components)
Slope	A number that describes both the direction and steepness of a line
Specificity	Ability of the method to assess, unequivocally, the analyte in the presence of other components that are expected to be present (e.g., impurities, degradation products, matrix components, etc.).
SD	The amount of variation or diversion of a set of values
Systematic error	A consistent and repeatable error typically indicating an issue with the assay or instrument
TAE	The amount of error that is allowed without invalidating the results. It includes both imprecision and inaccuracy in the method and accounts for both systematic and random errors. It is also referred to as total error, total analytical error, or allowable total error.
Within-run precision	Assesses the repeatability of the assay in the short term
Validation	A more comprehensive evaluation that establishes method performance by examining many characteristics of the assay. It allows the laboratory to evaluate the assay's performance in their patient population
Verification	A less comprehensive evaluation that confirms claims in manufacturers' package insert but does not confirm that the assay is suitable for the laboratories patient population

AMR, Analytical measurement range; *CLIA*, Clinical Laboratory Improvement Amendment; *CRR*, clinically reportable range; *CV*, coefficient of variation; *LOB*, limit of blank; *LOD*, limit of detection; *LOQ*, limit of quantification; *R*, correlation coefficient; *SD*, standard deviation; *TAE*, total allowable error.

challenging due to a potential lack of consensus among different clinicians and/or subspecialties. Importantly, analytical goals should be matrix-specific, particularly for testing available in multiple matrices. Once these goals are defined, only methods that meet these goals should be selected for evaluation and implementation.

Analytical goals provide the laboratory a benchmark to objectively evaluate data in order to help make a decision regarding method acceptability. If method selection is not goal-based, the laboratory runs the risk of relying on subjective data and/or assuming manufacturer claims are consistent with the needs of a target patient population. While important, other factors such as cost should not outweigh analytical performance. If laboratories have goals at the onset, in practice, it is less likely they will lose a fiscal argument typically posed by administration.

It is routine to define analytical goals as total allowable error (TAE). TAE is a simple and convenient parameter. It incorporates both imprecision and inaccuracy, or stated in other terms, random and systematic bias. The sum of these two sources of error should not exceed the analytical goal for TAE. Once the laboratory has defined analytical goals and other relevant testing parameters (e.g., cost and other logistics), they can initiate the method (or analytical platform) selection process. Prior to selecting an assay, homework should be done to narrow down options to a list of candidate assays and vendors that will likely meet the analytical goals. Published studies can help. In addition, laboratories may learn valuable information from attending pertinent academic conferences and vendor exhibitions, although they should be cognizant of relying on manufacturer claims without verification. Social media, or for those that remember rotary phones, a phone call to a colleague, can also assist with method selection.

The next decision after selecting a testing method is whether method verification or validation is warranted, as the two processes are different (Table 4.2). Method verification is a less comprehensive evaluation that confirms claims in manufacturers' package inserts, but verification does not confirm that the assay is suitable for a specific patient population. On the other hand, method validation is a more comprehensive assessment that establishes method performance by examining many characteristics of the assay. Method validation enables the laboratory to evaluate assay performance with respect to their laboratory infrastructure (i.e., instrumentation, workflow, and personnel) and their specific patient population. Laboratories should perform method verification prior to implementation. The remainder of this chapter will focus on method validation.

Helpful tips and tools prior to performing validation

Other preparatory steps may be helpful to complete prior to performing the validation. Purchasing a specialized

software program for evaluating and analyzing data for laboratory tests is highly recommended. Several options are available including EP Evaluator (Data Innovations, Burlington, VT, USA), Analyze-It (Analyze-it, Leeds, UK), or a homegrown template using R-programming packages [4]. Statistical evaluation of analytical parameters, such as precision, accuracy, reportable range, method comparison, sensitivity, and reference interval modules, is usually available in these programs. The laboratory should try to standardize and use one program throughout the laboratory, but ensure enough licenses for multiple users. Standardization will make it easier to train staff and review results.

Laboratory directors should review data from proficiency testing programs, such as CAP, as some proficiency tests are subdivided into specific techniques/instrumentation. This may inform the laboratory about possible bias between new and predicate methods, and thereby guide the selection of method validation experiments. For example, if the coefficient of variation (CV) is high at the low end of the analytical range, or if there is a bias at the high end of the analytical measuring range (AMR), the laboratory may want to include additional precision studies or additional correlations at the low and high ends of an assay, respectively.

Method vendors are frequently willing to assist with method validation, and we recommend requesting their input early in the process. Manufacturers may provide QC materials, calibrators, and reagents free of charge—it never hurts to ask. Some will even perform the experiments and enter the data into your software program of choice. Manufacturers should also be engaged in any troubleshooting during the method verification and/or validation to confirm the instrument and/or assay is configured properly. Note that if one is validating a third party product, manufacturer support may be challenging as manufacturers will typically only support vendor-specific reagents or products.

Laboratories should develop a validation plan. Table 4.3 summarizes the key validation components of unmodified nonwaived assays, as well as LDTs. Of note, it will often benefit the laboratory to save specimens with unusual or infrequently encountered results (e.g., positive hepatitis B IgM and high acetaminophen concentrations) for correlation studies (assuming the analyte is relatively stable at $-20^{\circ}C$ or $-80^{\circ}C$). It is also wise to begin with precision and reportable range (described in detail below) studies, which do not require patient samples. If the method does not meet the predefined precision and/or reportable specifications, one does not need to progress to finding patient specimens for method comparisons. The considerations for validation of a nonwaived FDA-approved assay and LDT differ (Table 4.1). In this chapter, we will first discuss method validation basics for FDA-approved assays, followed by the crucial components of LDT validation studies.

Method validation basics

Precision and accuracy

There are two popular ways to assess the imprecision of an assay: (1) within-run or repeatability; and (2) between-run or within-laboratory (Table 4.2) [5]. If multiple sites or instrumentation are involved, additional experiments, as outlined in the CLSI approved EP05-A3 (Evaluation of Precision of Quantitative Measurement Procedures) document, will be required [5]. In general, quantitative assays may include precision at the lower limit of quantification (LOQ), low, medium, and high ranges of the assay to characterize imprecision across the AMR. This is because, in many cases, imprecision is greater in magnitude as the analyte concentration approaches the LOQ. Limit of blank (LOB) and limit of detection (LOD) will be discussed later in this chapter.

Within-run precision ("repeatability") assesses assay precision over the short term, where variables such as reagent lot and operator are consistent. It typically represents the lowest imprecision the assay is able to achieve. Although matrix-specific patient samples are ideal, typically a laboratory will repeat QC samples at multiple analyte concentrations with 10−20 samples in one run in 1 day and calculate the mean, standard deviation (SD), and CV to determine assay imprecision. Practically speaking, if the repeatability of the assay is unacceptable, it is not worth further studies as it is unrealistic to achieve a better CV than that within-run, and an alternative assay or approach should be pursued by the laboratory.

Between-run or within-laboratory precision assesses the total imprecision of the assay. Most laboratories will repeat aliquoted QC at various levels two times for 10 consecutive days or one time for 20 consecutive days. Similar to within-run precision, the mean, SD, and CV are calculated. The recommended number of repeats and levels of QC varies. According to CLSI EP05-A3, a "20 × 2 × 2" design is ideal, where the laboratory runs two runs per day with two samples each run for 20 days (not necessarily consecutively) [5]. The levels should be spread evenly throughout the reportable range and include levels at or near the medical decision point(s). Patient pools can also be utilized instead of, or in addition, to QC material. A key benefit of using patient pools is that such samples will more realistically assess potential matrix effects. It is the responsibility of the medical director to decide whether or not to follow specific guidelines and which ones will be followed (e.g., CLSI and FDA).

Acceptable between-run imprecision is usually less than 25% of the TAE (Table 4.2). The level of imprecision

TABLE 4.3 Components of nonwaived method validation.

	Example studies	Important considerations
Unmodified nonwaived[a]		
Precision and accuracy	Within-run precision (repeatability):10–20 repeats at multiple levels in 1 day	If repeatability is not acceptable, troubleshooting should occur before any other studies as repeatability represents the best possible performance
	Between-run precision (total precision): multiple levels run once or twice per day for 20 days	Acceptability will vary by analyte
	Accuracy: compare results to reference materials or other materials with known concentrations	QC or proficiency samples can be utilized or comparison to a definitive or reference method can be performed
Reportable range	Three to seven specimens run multiple times with concentrations throughout the desired range. Assigned values are compared with measured values	Should span the medical decision point(s) and challenge the limits of the assay
Method comparison	Run a minimum of 40 specimens using the comparative and new method that span the medical decision point(s)	Unless the new method is compared with a reference or gold standard method, accuracy cannot be determined
Laboratory developed test (includes additional studies below)[a]		
Carryover	Run specific sequence of high and low QC samples in one continuous experiment	Acceptability will depend on clinical context
Matrix effects	Run at least three QC specimens from at least two analyte concentrations each; compare specimens prepared in independent sources (preferred) or pooled assay matrix with those prepared in solvent	Acceptability will depend on clinical context; can also perform postcolumn infusion for chromatographic techniques
Stability	Run at least three QC specimens from at least two concentrations each for short-term, long-term, and postprocessing stability	Acceptability will depend on clinical context; stability conditions (temperature, duration, freeze–thaw, etc.) are determined by assay workflow and expected handling/storage conditions
Selectivity/ specificity	Assess at least six independent blank biological matrices for otherwise false-positive results. Spike analyte-free matrix with known analyte analogs in at least two QC levels	Run each analyte separately for multiplex assays
Interferences	Spike known interferants (exogenous and endogenous) in at least two QC levels	Acceptability will depend on clinical context
Recovery	Run at least three QC specimens from at least two analyte concentrations each; compare spiked specimens extracted from pooled matrix with blanks extracted then spiked with identical levels of analytes	Recovery of less than 100% indicates incomplete extraction
Sensitivity	Run at least three calibration experiments along with blank specimens; determine SD of calibration slope and blank response	Multiple methodologies to estimate sensitivity
Dilution	Within-range: for each dilution factor, dilute a high QC range specimen with five technical replicates. Above-range: for each dilution scheme, dilute an above-ULOQ specimen with five technical replicates	For above-range dilutions, calculate theoretical recovery
Specimen requirements	Determine acceptable matrices, containers, handling, and storage conditions	Base handling and storage conditions on stability validation

LOQ, Upper limit of quantitation; *QC*, quality control; *SD*, standard deviation.
[a]*For specific numbers of replicates, analyte concentrations, and other experimental design factors, users should consult specific guidelines and established protocols (e.g., CLSI).*

varies by test and methodology, with the majority of routine automated chemistry analytes, such as sodium and potassium, exhibiting imprecision far less than 10% and LDTs having imprecision between 10% and 25%. Unacceptable precision can be due to a variety of factors. Poor repeatability (within-run) may be due to variation in pipetted volumes, whereas poor between-run precision may be due to reagent lot changes, calibration cycles, and/or operator rotation. Before rejecting the studies, the laboratory should seek to determine the cause of any high imprecision.

Precision data cannot only determine assay imprecision, but it may also provide information on QC or instrument performance. Trends in QC or patient pool results (whichever are used for precision studies) over time should be reviewed. If QC results continue to drift over time, it is concerning and may indicate a degradation in QC materials or an instrumentation failure (Fig. 4.1). If there is an issue with analyte stability, results of patient pools may decrease over time. Moreover, if results continue to rise over time, it may indicate loss of calibration. Concerning results should be reviewed with the manufacturer.

Samples with known concentrations, such as QC materials, can be utilized to establish accuracy; "the ability to recover the correct amount of analyte present in a specimen" (Table 4.2). Laboratories can also include measurement of remnant or previously evaluated proficiency testing samples in their method validation studies to verify accuracy. Although it can be more challenging, comparison of the method or a subset of samples with a definitive or reference method is ideal to ensure that results are accurate.

Reportable range

According to CLIA, reportable range is defined as "the span of test result values over which the laboratory can establish or verify the accuracy of the instrument or test system measurement response" [3,6]. The terms linearity, assay range, analytical range, AMR, and clinically reportable range (CRR) may also be used, and are defined in Table 4.2. While AMR is defined as the range of values an assay can quantitatively report without additional sample dilution or concentration, CRR refers to the range of values an assay can quantitatively report with sample pretreatment such as dilution or concentration.

Linearity assesses the relationship between observed values and a known concentration. It is important for clinicians to know that the relationship between the known concentration and the reported result is linear as they use results to make medical decisions. CLSI guideline EP06-A provides recommendations on how to assess linearity [5]. Laboratories can purchase linearity material or dilute high-concentration patient pools. If it is challenging to find high concentrations of some analytes, such as therapeutic drugs, the laboratory may need to spike samples using high-concentration reference materials. Importantly, matrix-appropriate specimens should be diluted according to the manufacturers' instructions (e.g., with manufacture diluent). The number of samples utilized to determine the reportable range for a nonwaived FDA-approved assay varies from three to five; five to seven independent concentrations are recommended according to CLSI EP-06A [6]. Regardless of the guidance followed, the laboratory should ensure that the samples span the medically relevant range as well as challenge the limits of the assay (i.e., choose concentrations expected to fail). The number of repeats of the material also varies from three to four and provides additional assessment of imprecision in the method.

Software programs are frequently utilized to display the reportable range with assigned values on the *x*-axis and reported values on the *y*-axis. Recovery is defined as 100 times (×) the measured mean of the replicates divided by the assigned value (Table 4.2). Visual inspection of the recovery and comparison of assigned and reported values are important. Outliers or concerning trends can usually be detected after visual inspection alone. However, recent CLSI guidelines (EP06-A) suggest that the polynomial method should be utilized to analyze the data [6]. This method estimates the magnitude of nonlinearity at every level and controls for unacceptable repeatability. Many programs will perform the polynomial analysis, but acceptability criteria are user-defined and based on medical significance.

Individual points should be plotted as opposed to average concentrations, which can conceal imprecision. Troubleshooting is required if the plot reveals deviations from linearity or evidence of instrument failure. The laboratory may need to involve the manufacturer to assist with troubleshooting. Common reasons for nonlinearity include improper sample preparation, an interference, or lack of calibration. Ultimately, if the reported value deviates significantly from the assigned value and cannot be corrected, the level should not pass and the preceding level should be used to define the upper limit of the reportable range. If results above the upper limit are medically required, laboratories may dilute samples into range and report results if dilutional challenges meet acceptance criteria.

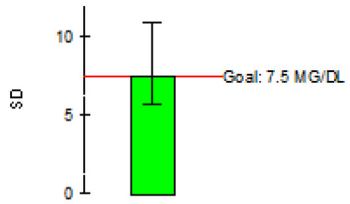
Method comparison

Method comparison is the testing of the same sample by at least two independent laboratory assays for the same analyte for the purpose of characterizing test agreement including systemic and random errors. Laboratories should ideally compare their new method to a reference or gold standard method. However, this can be impractical, and consequently, many laboratories often compare

IgM
Instrument Instrument 1
Sample Name Plasma

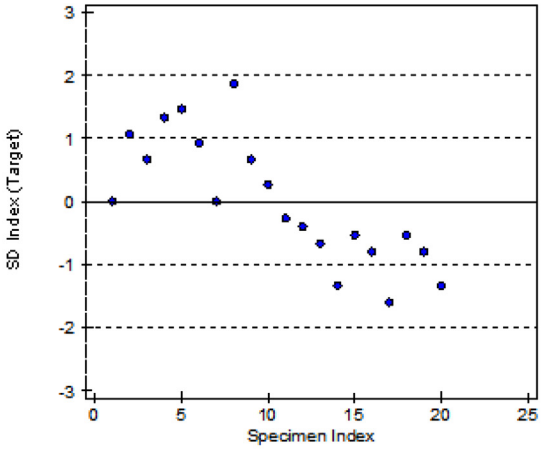
Simple Precision

Precision Statistics



Obs Standard Dev (SD)	7.5
Pass/Fail/Uncertain	Yes
95% Confidence for Obs SD	5.7 to 11.0
Obs Coef of Variation (CV)	10.0%
Obs Mean	75.0 MG/DL
Number of Specimens (N)	20 of 20
95% CI for Obs Mean	71.5 to 78.5
Obs 2 SD Range	60.0 to 90.0

Precision Plot



User's Specifications

Precision Verification Goal	TEa
Allowable Total Error	75.0 MG/DL (conc) or 20.0%
Random Error Budget	10%
Allowable Random Error	7.5 MG/DL (conc) or 2.0%

Supporting Data

Analyst	Riley
Expt Date	03 Jul 2018
Units	MG/DL
Target Mean	75
Target CV	10.0
Comment	

Precision Data

Index	Result	Index	Result	Index	Result	Index	Result
1	75	6	82	11	73	16	69
2	83	7	75	12	72	17	63
3	80	8	89	13	70	18	71
4	85	9	80	14	65	19	69
5	86	10	77	15	71	20	65

X: excluded from calculations

FIGURE 4.1 Precision experiment. Quality control with a target mean of 75 mg/dL of IgM was run over 20 days. The coefficient of variation, standard deviation, and 2 SD range were 10%, 7.5, and 60.0–90.0, respectively. The experiment passes; however, there is a concerning downward trend in results over time that should be investigated.

the new method with their existing method. A major limitation to this approach is that the laboratory cannot determine if the results are accurate; only if a clinically significant bias exists, that will affect patient care.

There are two important factors to consider when comparing methods: (1) the number of specimens in the comparison evaluation; and (2) the range of analyte concentrations or measurements in the comparison specimens. Both CLIA and CLSI have recommended a minimum of 40 specimens to be included in the comparative analysis [3,7–10]. Just as important as the quantity of samples, the specimens should span the medical

decision points and allow for an accurate assessment of bias across the AMR, and possibly even the CRR. A comparison of 40 specimens clustered around the AMR median may falsely reassure the laboratory and hide a bias at the low or high end of the AMR. The method comparison study should preferably occur over several days to reveal drift or change in results, though care should be taken to ensure specimen stability during the entire comparison.

Least squares linear regression ($y = mx + b$; also referred to as $Y = a + bX$ in other texts), either Deming or Passing–Bablok, can be utilized to compare results and detect any bias (Table 4.2). Most software programs allow the laboratory to choose their preferred analysis. X is the comparative method (plotted on the x-axis), y is the new method (plotted on the y-axis), m is the slope, and b is the y-intercept. Both Deming and Passing–Bablok regression models reasonably assume that each method is subject to error. Deming regression is usually the preferred analysis if results are well distributed and outliers are not present. Deming regression is also used if medical decision points are being compared. On the other hand, Passing–Bablok regression is preferred if outliers are present, as the Passing–Bablok regression model employs a nonparametric regression, which means the analysis is constructed by using the data as opposed to predetermined parameters. Outliers should not be removed from the analysis unless the cause is known (e.g., pipetting error). The cause of outliers should be determined, and only if appropriate may be removed from data analysis. Of note, Passing–Bablok regression analysis should not be used if the comparison has a very low or high number of specimens, or if a comparison of medical decision points is the primary goal.

Results can be presented in different formats to assist with interpretation. Analysis of the regression line is usually the first step. The correlation coefficient (R) indicates the degree of correlation and adequate range of data, but does not indicate how the methods compare. If R is less than 0.95 (some laboratories may set a looser limit of 0.90 or a stricter limit of 0.975), the laboratory needs to assess the range of data. The slope, y-intercept, and confidence intervals are also important. Identical methods have a slope equal to 1.0 and an intercept of zero at the 95% confidence interval; a feat rarely achieved.

Laboratories must also review method comparison data for bias. If the results of the new method are consistently higher or lower than the current method, a constant bias is present. The intercept represented by b in the straight line equation $y = mx + b$ represents the constant bias (Fig. 4.2). Bias or Bland–Altman plots are most commonly employed to investigate proportional bias, where results are always higher or lower by the same proportion. Proportional bias is represented by slope or the m in the equation of a straight line ($y = m\,x + b$) (Fig. 4.3). In some cases, the laboratory will anticipate a bias such as when switching to a different immunoassay (e.g., parathyroid hormone and thyroid stimulating hormone) or

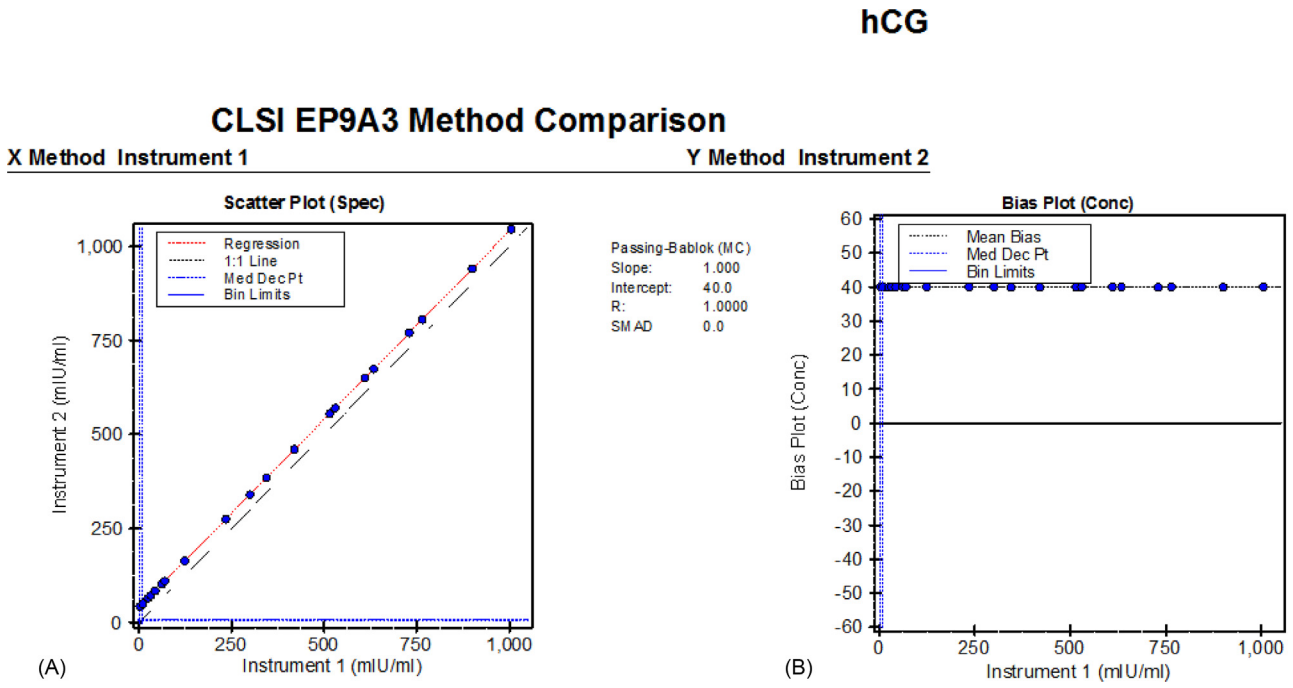


FIGURE 4.2 Constant bias. A method comparison experiment for human chorionic gonadotropin was performed. The Passing–Bablok regression analysis shows a constant bias of 40 mIU/mL (i.e., intercept of 40.0) between the two methods with a slope and R of 1.00. (A) Scatter plot. (B) Bias plot.

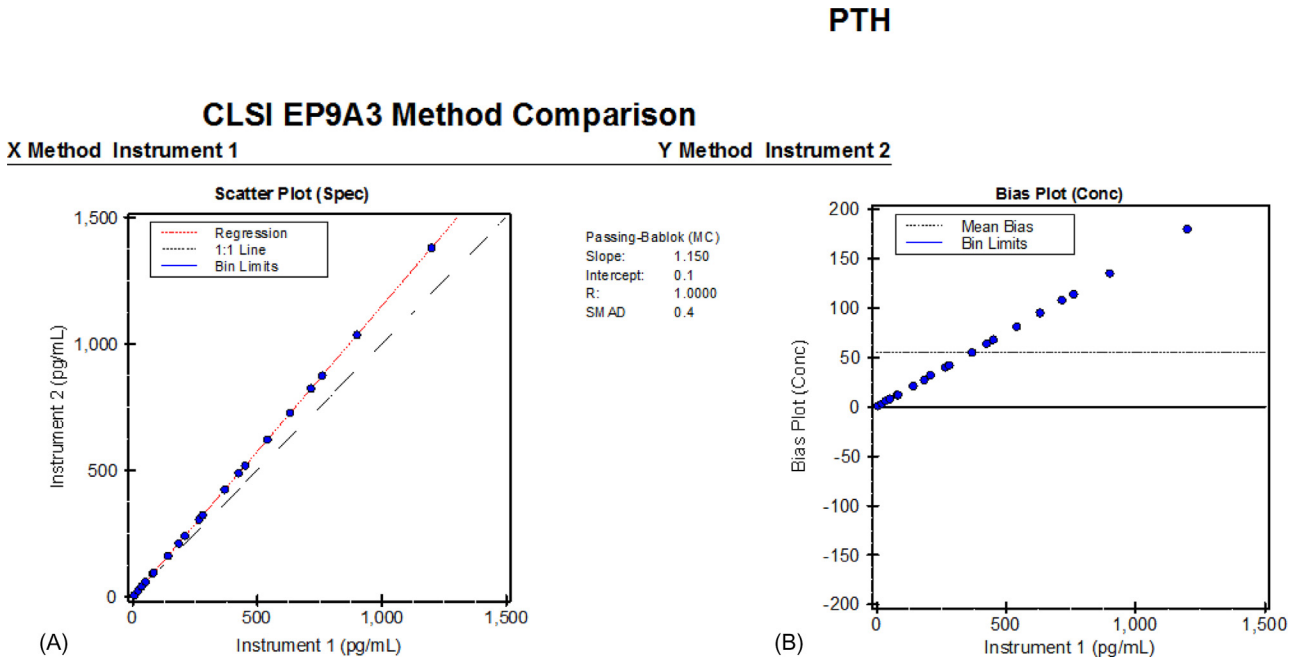


FIGURE 4.3 Proportional bias. A method comparison experiment for PTH was performed. The Passing−Bablok regression analysis shows a proportional bias of 15% (i.e., slope of 1.15) with an intercept of 0.1 and an *R* of 1.00. (A) Scatter plot. (B) Bias plot. *PTH*, parathyroid hormone.

changing methodologies (e.g., calcium and lipids). Proficiency testing results can be helpful to anticipate or confirm the type of bias. A clinical review of the data is critical to determine whether the bias is acceptable and if the reference range should be adjusted.

Medical decision point analysis is frequently employed for assays that have standard cutoffs or reference ranges such as therapeutic drugs, fasting glucose concentrations, and lipid profile measurements. The laboratory should calculate the percentage of cases for which the medical decision (e.g., both results higher than the reference range or cutoff) would be the same for both methods. Special consideration to unique populations should be given during method comparison studies. In addition, laboratories may need to verify how a method performs in pediatric patients, pregnant patients, or the critically ill; the latter of which may be difficult to define. If possible, the laboratory should gather specimens from these specialized populations and perform a focused method comparison.

A detailed discussion on reference intervals can be found in Chapter 3: Reference intervals: theory and practice. At a minimum, laboratories should perform a reference range verification study in order to verify the range provided by the manufacturer is acceptable in their patient population. To accomplish this, a laboratory can perform testing on 20 healthy individuals. If 18 out of 20 (90%) fall within the reference range, the reference range is considered verified. For some tests, a therapeutic interval is more appropriate than a reference interval (e.g., therapeutic drugs); therefore the laboratory may utilize published ranges believed to represent their patient population.

Validation of laboratory developed tests

As defined by the FDA, an LDT is "a type of in vitro diagnostic test that is designed, manufactured, and used within a single laboratory." These LDTs can be used to measure a variety of analytes, including ions, small molecules (xenobiotics and metabolites), and biological macromolecules (peptides, proteins, and nucleic acids), in a variety of matrices (blood, tissue, urine, saliva, and miscellaneous body fluids). LDTs enable laboratories to create customized assays to meet specific patient and system needs, such as for a specific patient population or if a given test is not commercially available. For example, a tertiary cancer center with a high volume of lymphoplasmacytic lymphoma ("Waldenstrom's macroglobulinemia") may choose to develop and validate an *MYD88* variant assay [11], whereas such a test may not be suitable for a local clinic or a rural hospital laboratory.

As they are not approved by the FDA, the performance of an LDT should be characterized for multiple variables by the performing laboratory. Assay parameters include the aforementioned precision, accuracy, reportable range, and method comparison experiments, as well as matrix effects, carryover, analyte stability, selectivity, interferences, dilution, sample preparation, and reagent stability. In addition to clinical context, the LDT validation must also take into account the specific assay

technology. For example, chromatography-based assays should rigorously characterize carryover, while ligand-binding assays should evaluate dilution scheme for possible prozone effects ("Hook effects"). Acceptability guidelines may also differ slightly depending on the assay technology, as current FDA bioanalytical guidelines do so for chromatography- and ligand-binding assays [12].

Carryover validation

Carryover is the unintended transfer of substances (analyte, reagent, and diluent) between samples or assay compartments (Table 4.2). This transfer can then perturb the readout of results for subsequent samples(s) [13]. While carryover usually reflects an immediately preceding or neighboring sample, it can appear to occur randomly and may reflect a delayed elution from a nonconsecutive sample. Carryover can occur anywhere along the assay procedure, including instrumentation, liquid handlers, and consumables.

Carryover is pertinent to a variety of assay methodologies, from automated chemistry analyzers to chromatographic methods to molecular diagnostics. Carryover is rather intuitive for chromatography-based methods such as ultra/high-performance liquid chromatography–mass spectrometry (U/HPLC-MS) [14]. For example, a sample may be retained on an autosampler or chromatographic column with suboptimal washing and reequilibration, only to be eluted and detected in the next sample. In automated chemistry analyzers and autosamplers, carryover can occur when cuvettes, sampling probes, or liquid handling components are contaminated or incompletely washed between samples [15,16]. For molecular assays, carryover usually refers to sample contamination from nucleic acids from environmental sources, previous samples, or concurrent samples [17]. For polymerase chain reaction (PCR)-based assays, even minor carryover or contamination can be significant because of analyte amplification.

Carryover can often be mitigated, though specific approaches will vary depending on the analytical methodology and source of carryover. For chromatography-based systems, this can be accomplished with alternative chromatography conditions, such as different columns or mobile phase conditions, longer system washes, multiple wash cycles, or longer equilibration times, even though such modifications may be at the expense of throughput [14,18]. When carryover is unavoidable, standard operating procedures (SOPs) can dictate running blank samples immediately after extraordinarily high analyte concentrations before running a subsequent sample. For molecular assays, carryover/contamination can be mitigated by careful laboratory design (separate pre- and postamplification areas), sterilization, automation, and certain reagents such as dUTP with uracil DNA glycosylase [19].

Carryover is formally evaluated by repeated testing of a high and low QC in a specific order [20]. One usually defines an acceptable level of carryover prospectively, based on factors such as clinical context and analytical performance. FDA bioanalytical guidelines recommend characterizing carryover, and using an acceptability cutoff of less than 20% lower limit of quantitation (LLOQ) [12]. Alternatively, some laboratories opt to use LLOQ and (upper limit of quantitation) ULOQ calibrators for carryover analyses as well. One suggested experiment is to run consecutive low (L) and high (H) QC samples in the following order: L1, L2, L3, H1, H2, L4, H3, H4, L5, L6, L7, L8, H5, H6, L9, H7, H8, L10, H9, H10, and L11. Carryover is assessed by looking for changes in the low QC samples, where multiple low QC samples are initially performed to define assay imprecision. Carryover would be expected in low QC samples immediately following high QC samples. Analysis can be aided by specialized software. For more extensive discussions of carryover, we refer the reader to CLSI protocol EP10-A3 [8].

Stability validation

The stability of an analyte during testing is a prerequisite for accurate and precise measurement. Certain analytes are susceptible to various forms of instability, which can be caused by chemical instability (e.g., nonenzymatic hydrolysis, oxidation, proteolysis, and photosensitive reactions), insolubility (e.g., precipitation or aggregation), protein denaturation (i.e., for enzymes), and contamination (i.e., live-cell samples). Characterization of analyte stability enables laboratory directors to select optimal handling and storage conditions to minimize the chance that a particular result is inaccurate.

Stability is relatively straightforward to assess (Table 4.4). Percent recovery in stored samples are calculated relative to control samples that are either unperturbed (e.g., not freeze–thawed or processed) or prepared fresh at the time of experiment (i.e., for short- and long-term storage assessments). Stability sample results should generally be within 15% of nominal concentrations (may vary, and FDA bioanalytical guidelines allow for up to 20% difference from nominal for ligand-binding assays). For quantitative assays, experiments should evaluate stability at two analyte concentrations (L and H) with at least three technical replicates to assess precision. Ideally, stability studies should utilize storage containers to be used in clinical operations. Alternatively, the percent difference (%DIF) between the stored or treated sample and the control sample can be calculated to evaluate acceptable stability. Even though the FDA recommends a comparison to the

TABLE 4.4 Components of stability validation.

Ambient conditions	Room temperature stability, usually a minimum of 6 h but often 12–24 h to model bench-top handling
Postprocessing	(i.e., after all reagents and sample preparation have been performed) should be characterized to determine the length of time a processed sample is expected to perform within acceptable performance specifications; in many cases, this involves testing a sample stored in an instrument autosampler to mimic a prolonged sample queue
Short-term storage	Refrigeration (4°C), often up to several weeks, to model typical laboratory workflows
Longer-term storage	Freezing (−20°C and/or −80°C), often up to several months depending on anticipated storage duration
Freeze–thaw	Samples should be tested for stability after a minimum of three complete freeze–thaw cycles, as certain analytes irreversibly precipitate or denature; in most scenarios, freezing should be done overnight and thawing should be performed at ambient temperatures

nominal concentration, calculating the %DIF excludes any bias introduced during the analyte spiking process.

Stability can be calculated as percent deviation from a theoretical (nominal) concentration (%NOM), or alternatively, it is calculated as a %DIF between stability-challenged and fresh (nonchallenged) QC samples. Both FDA and its European counterpart, the European Medicines Agency, discuss stability in terms of nominal concentrations [12]. In practice, percent nominal concentrations are determined by measuring samples immediately after their applied stability challenge, and by calculating the percent deviation from the nominal concentration determined by a freshly prepared calibration curve. By contrast, %DIF measures concentrations of the unperturbed and the stability-challenged samples in the same analytical run. The advantage of the latter approach is that each sample type is subjected to the same analytical run conditions.

Analyte stability should be characterized across the expected clinical sample handling and storage conditions, from sample collection to specimen shipment, storage, processing, and testing [21,22]. This typically includes ambient conditions, short- and long-term storage, freeze-thaw, and postprocessing stability.

In addition, the stability of analyte and internal standard stock solutions should be characterized, especially when the stock solutions are stored in different physical states or buffers from the source solution. Notably, in clinical operations, samples and reagents should not be utilized, which have been stored outside of the validated stability conditions. For more extensive discussions of reagent stability testing, we refer the reader to CLSI protocol EP25-A [23].

Selectivity/specificity and interference validation

Selectivity (often incorrectly used interchangeably with *specificity*) is defined by the FDA as "the ability of the bioanalytical method to measure and differentiate the analytes in the presence of components that may be expected to be present (Table 4.2). These could include metabolites, impurities, degradants, or matrix components." A similar concept is specificity, defined by the FDA as "the ability of the method to assess, unequivocally, the analyte in the presence of other components that are expected to be present (e.g., impurities, degradation products, matrix components, etc.)" [12]. In practice, it can be extraordinarily difficult to ascertain *unequivocal* specificity; therefore the former term selectivity is often preferred.

Each LDT should therefore assess the assay readout for analyte selectivity. In other words, this means providing evidence that the readout is *actually* quantifying the intended analyte. In principle, biological samples may contain nonanalyte substances that result in appreciable readouts. For example, analyte isoforms, isozymes, or posttranslationally modified analytes may contain epitopes recognized by a reagent antibody in an immunoassay. The FDA therefore recommends evaluating six independent sources of the relevant blank biological matrix to determine if endogenous substances result in significant signal in the analyte readout LLOQ (n.b., 10 sources for ligand-binding assays).

Interference is the modulation of the assay readout by nonanalyte substances in a sample, often referred to as "artifacts." There are two main types of interferences: endogenous and exogenous [24–26]. Exogenous interferences originate from outside the body and are therefore not normally found in clinical specimens. Each type of interferent can cause positive or negative bias in analytical results, depending on the assay methodology and other factors such as interferent concentration. In this way, sources of interference work to effectively reduce the selectivity and specificity of an LDT.

Examples of common potential endogenous interfering substances include bilirubin (i.e., icteric samples) and lipids (e.g., hypertriglyceridemia), which can interfere with light-based readouts; hemolysis, via release of red blood cell intracellular contents such as potassium, abundant enzymes like LDH and AST, and hemoglobin; paraproteins; immunoglobulins; and endogenous substances

chemically similar to target analytes of interest. For endogenous biomarkers, endogenous interfering substances may also include macromolecules with homology to the analyte of interest (proproteins/enzymes and isoforms). For some biological matrices, such as urine, normal biological variation can effectively create interferences in the form of protons (pH) and salts. Endogenous interfering substances can manifest as matrix effects.

For ligand-binding assays measuring endogenous analytes, an important related concept is parallelism (Table 4.2). The FDA notes, "Parallelism demonstrates that the serially diluted incurred sample response curve is parallel to the calibration curve." Why perform this experiment? The rationale is to verify if the binding characteristic of the endogenous ligand to the binding agent (e.g., antigen−antibody) is the same as for the calibrator [27]. While similar, this is not to be confused with dilutional studies. Parallelism validation utilizes samples containing high levels (near but less than the ULOQ) of an endogenous analyte, which are diluted to assess whether the dilution−response curve is parallel to a standard concentration−response curve [28]. Parallelism can flag potential matrix effects and interactions between critical assay reagents. There are no established criteria for what constitutes acceptable parallelism, but in general, test samples should be within 20%−30% of the corresponding calibrators [29−31].

Examples of potential exogenous interfering substances include heavy metals, where conventional urinalysis containers may leach trace amounts of heavy metals into otherwise analyte-free sample [32]; xenobiotics (e.g., other medications) and metabolites, where certain related and unrelated drugs may cross-react with immunoassay-based therapeutic drug monitoring assays or therapeutic monoclonal antibodies may produce positive readouts on protein electrophoresis [33]; adulterants, where oxidants, acids, or bases may be added to samples in illicit urine drug testing to chemically modify the tested analyte; and postcollection analyte addition, where urine samples spiked with drug after voiding to simulate compliance [34].

In mass spectrometry (MS), isobaric but chemically distinct substances may coelute with the assay analyte. As an example, consider a U/HPLC-MS/MS assay to quantify morphine-3β-glucuronide (M3βG). Since normal human metabolism also creates morphine-6β-glucuronide (M6βG), a selective method must distinguish between these two isobaric analytes so as not to inaccurately quantify M6βG as M3βG. In multiplexed methods, selectivity validation includes testing each analyte individually and measuring the readout of each multiplexed analyte (Fig. 4.4).

An LDT should characterize the effects of expected sample interferences, which is often done by determining the percent recovery of a spiked sample containing a potential interferent with a control sample. Interference can be tested in two manners: (1) in blank matrix to evaluate for signal creation; and (2) in QC or analyte-spiked samples to evaluate signal attenuation. In most clinical assay formats, the former would lead to false-positive readouts, while the latter would produce false-negative readouts.

Endogenous interfering substances can be assessed by several approaches [35,36]. First, a laboratory can utilize clinical samples of suboptimal specimen integrity, such as grossly icteric, lipemic, or hemolyzed samples. These samples can be tested as is, and also mixed with a known amount of analyte to compare percent recovery with a noncompromised sample containing the same amount of the analyte of interest. The advantages of this approach are the relative ease of acquiring samples and straightforward procedure, though the main disadvantage is the lack of standardization with respect to interferants from sample to sample. Therefore when using this approach, attempts should be made to quantify the interferant, such as

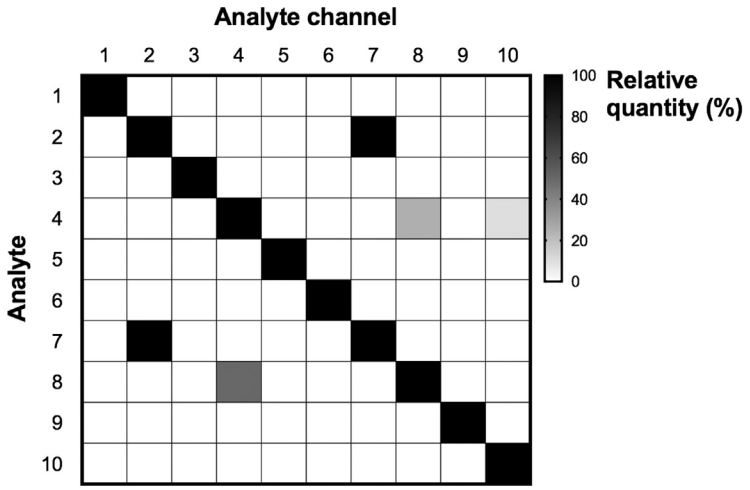


FIGURE 4.4 Selectivity testing in a multiplex laboratory developed test validation. In this hypothetical assay, 100 units of each analyte (1−10) are analyzed by each analyte channel (1−10). Analytes 2 and 7 are quantified at the same concentration, suggesting the method cannot differentiate between the two analytes. Analytes 4 and 8 cannot be differentiated, though the method does not measure the same quantities unlike analytes 2 and 7. Analyte 4 is identified in analyte channel 10, which could be due a sample impurity.

measuring hemolysis or icteric indices. Note that these indices are semiquantitative measurements based on light absorbances, and not known interfering substance concentrations.

Another approach to evaluate endogenous interfering substances is to spike blank matrix and/or QC samples with pure interferents, and compare percent recovery to noncompromised controls. This approach is more rigorous, though certain purified analytes are relatively expensive or unavailable, and such prepared samples may not fully create the biological milieu needed for interference. Whenever practical, the interference substance to be spiked should be in concentrated form (e.g., 20X). The purpose of using concentrates is to prevent significant dilution of the sample matrix and analytes upon spiking. This better enables the assessment of the actual interfering substances on assay results by not significantly diluting the sample matrix (especially in cases where the diluent may not fully approximate the sample matrix) and by allowing higher analyte concentrations (by not significantly diluting the QC samples).

Exogenous interferent testing is relatively straightforward, and usually involves spiking blank matrix and/or QC samples with known amounts of the exogenous interferent. Exogenous interferents could include concomitant medications in the patient population [37]. As there are no strict guidelines, specific interfering substance tolerabilities will depend on the discretion of the medical director, which involves evaluating results within the appropriate clinical context. For more extensive discussions of interferences, we refer the reader to CLSI protocol EP07-A3 [38].

Matrix effects validation

Matrix effects are the direct or indirect alteration in response due to the presence of unintended analytes or other interfering substances in a sample; in many ways, matrix effects studies may be viewed as a subset of interference validation. Endogenous matrix components include salts, carbohydrates, amines, lipids, peptides, nucleic acids, and related metabolites. Exogenous matrix components include additives from sample collection (e.g., heparin and calcium) or reagent components. The nature of matrix effects depends on the assay analyte, testing methodology, and matrix itself, and is often unpredictable a priori. For MS-based systems, matrix effects are typically manifested as ion suppression, though ion enhancement can occur as well [39].

There is actually no mandated threshold for matrix effects, meaning the degree of acceptable matrix effects are decisions by the medical director to be based on a variety of factors (e.g., clinical context and expended incidence). At a minimum, matrix effects must be characterized as part of an LDT validation [40]. This should include evaluation of at least two analyte concentrations, as the magnitude of matrix effects may be analyte concentration-dependent. There are several experimental approaches to characterize matrix effects, depending on the LDT methodology. For MS-based assays, this may include (1) constant postcolumn infusion of an analyte with a matrix injected into the chromatographic system; or (2) spiking different matrices in pre- and postextracted samples and comparing percent recovery [41,42]. Both can be relatively time-consuming, and the former is qualitative in nature, whereas the latter is quantitative. Postcolumn infusion examines matrix effects across the entire chromatographic run, whereas pre- and postextraction addition examines matrix effects at the retention time of the analyte of interest. Furthermore, postextraction addition is most appropriate when using stable-isotope internal standards. A more general method for assessing potential matrix effects is to spike at least 10 independent biological samples and assess variability (SD and CV). An important caveat to this general approach is that it reflects the mean matrix effects, and may not reflect matrix effects from individual samples [43].

There are several strategies to mitigate the effects of matrix effects in LDTs. A general strategy to mitigate matrix effects is sample dilution, which can potentially decrease the effective concentration of the interfering substance(s) in the sample and presumably its effect on the assay readout [44]. The amount of sample dilution must balance analytical sensitivity requirements and potential imprecision and bias introduced from dilution. If performed clinically, sample dilution constitutes a necessary validation parameter. Another strategy to mitigate matrix effects includes sample purification such as solid-phase extraction, liquid−liquid extraction, protein precipitation, or immunoenrichment, which can reduce the relative concentration of interfering matrix components. In assays involving chromatography, various parameters such as solvent composition and gradient can also be optimized. When measuring small molecules by MS, matrix effects are often mitigated through the use of deuterated internal standards [44]. In practice, such compounds may be unavailable or prohibitively expensive, forcing laboratorians to utilize internal standards with similar chemical properties and assay behaviors confirmed through experimental validation.

Recovery validation

The FDA defines recovery of an analyte as “the detector response obtained from an amount of the analyte added to and extracted from the biological matrix, compared with the detector response obtained for the true concentration of the analyte in solvent.” Recovery is especially important for analytes requiring preanalytical extraction,

purification, and/or enrichment. For evaluating recovery in the context of extraction, matrices are spiked with analyte, and then extracted. This value is then compared with blank matrices spiked with the same amount of analyte after undergoing an otherwise identical extraction. Recovery is then calculated by the relationship:

$$\%\text{Recovery} = [\text{Analyte}]_{\text{spiked}\rightarrow\text{extracted}} - [\text{Analyte}]_{\text{blank extracted}\rightarrow\text{spiked}} / [\text{Analyte}]_{\text{blank extracted}\rightarrow\text{spiked}} * 100$$

Notably, analyte recovery does not have to be 100%. However, analyte and corresponding internal standard recovery should be characterized for precision and reproducibility at multiple analyte concentrations (e.g., low, medium, and high QC values) by comparing measurements with control samples not subjected to extraction.

Sensitivity validation

The analytical sensitivity of an LDT should be characterized to determine the lowest concentration of analyte that can be reliably detected above the expected sample background. Some key parameters related to sensitivity are LOB, LOD, and limit of quantification (LOQ) (Fig. 4.5 and Table 4.2) [45]. The LOB is the highest apparent analyte concentration expected in blank sample replicates containing no analyte [$\text{LOB} = \text{mean}_{\text{blank}} + 1.645\ (\text{SD}_{\text{blank}})$], while the LOD is the lowest analyte concentration likely to be reliably distinguished from the LOB and at which detection is feasible.

There are a variety of methods that can estimate the LOD that incorporate various statistical assumptions. Ultimately, the best choice of estimating LOD depends on the LDT clinical context [46–48]. Often, the LOD can be determined by utilizing both the measured LOB and test replicates of a sample known to contain a low concentration of analyte [e.g., $\text{LOD} = \text{LOB} + 1.645\ (\text{SD}_{\text{low concentration sample}})$]. Other protocols can define LOD in terms of signal:noise ratios (e.g., greater than 3:1 or 10:1 signal:noise) and various multipliers based on imprecision. Another approach involves empirically testing serial dilutions of analytes until the analyte cannot be reliably detected. One practical approach to estimating LOD involves measuring repeated blank samples for imprecision, and calculating LOD based on the calibration curve using the relationship: $\text{LOD} = 3.3\ \text{SD}_{\text{Blank}}$/calibration curve slope [49].

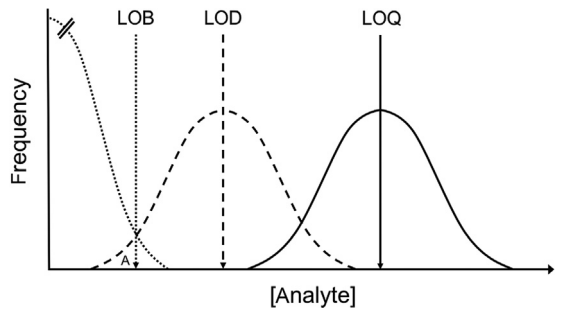


FIGURE 4.5 Key parameters related to assay sensitivity. The LOQ (solid line) and LOD (*dashed line*) are chosen, such that they meet the performance specifications for imprecision and bias. The LOD is chosen, such that a small proportion of samples with analytes at the LOD samples falls below the LOB (*dotted line*; "A"), and similarly, only a small/essentially nonzero proportion of blank samples exceeds the LOD. *LOB*, limit of blank; *LOD*, limit of detection; *LOQ*, lower limit of quantification.

The LOQ is the lowest concentration at which the analyte cannot only be reliably detected but at which some predefined goals for bias and imprecision are met. Note that the LOQ may be equivalent to LOD or it could be at a much higher concentration. Laboratories may also set the LOQ higher than the true LOQ depending on the clinical context (e.g., to avoid reporting analytically positive yet clinically insignificant findings).

Dilution validation

Dilution is the addition of a diluent (e.g., water, pooled serum, and buffer) to a sample. Sample dilution is often used when the analyte concentration is above the ULOQ or if there is a suspected interferent. An LDT should characterize the effects of sample dilution, specifically for percent recovery. In a well-behaved (linear) system, the amount of analyte detected in a diluted sample should be inversely proportional to the dilution factor. In certain cases, an analyte may be protein-bound and will not achieve equilibrium under the diluted conditions. For ligand-binding assays, dilution studies are also performed to rule out prozone ("Hook") effects, whereby exceedingly high analyte:ligand ratios disrupt expected binding behaviors [50].

The main considerations for dilution validation are the choice of diluent, the magnitude of dilutions, and testing within- and outside-range dilutions [51]. One should also consider the choice of diluent to be compatible with the testing methodology, and in general, "like" matrices should be utilized. For example, assays normally run in serum may dilute with pooled negative human serum. If not feasible (e.g., cases of rare matrices), a buffered solution or water may suffice. The magnitude of dilution should be based on the expected analyte concentrations. For example, if samples are occasionally encountered up to fivefold the ULOQ, a proposed dilution validation could examine the effect of 5- and 10-fold dilutions. Each dilution, whether it is within range or outside range, should be evaluated as a series of multiple technical replicates (e.g., five per dilution level), with acceptable precision between 15% and 20% CV and acceptable accuracy between 15% and 20% nominal

concentrations. Additional suggestions for dilution validation schemes can be found in CLSI documentation (e.g., EP34 and C62-A).

Reference standards

Assay performance is routinely compared with calibrators and QC samples derived from reference standards, which also include internal standards. Impure reference standards can adversely affect assay performance, most notably accuracy. Therefore it is critical to utilize reference standards with rigorously characterized identity and purity, and traceable lines of production and storage. For example, in therapeutic drug monitoring, a reference drug should have its chemical identity verified by multiple techniques such as ^{1}H nuclear magnetic resonance, MS, and physical properties. The same applies for redundant assays for analyte purity such as U/HPLC-MS, elemental analysis, inorganic content analysis, and residual solvent analysis. Other analytes such as proteins and nucleic acids should be characterized with appropriate sequencing technologies. Whenever possible, chemical identities identical to the biological analyte should be utilized.

Reference standards are usually derived from three sources: certified reference standards (e.g., USP compendium), commercial-supplied reference standards from reputable sources (e.g., Cerilliant, LGC), and custom-made reference standards. Regardless of the source, LDTs should carefully catalog the lot, batch number, expiration date, certificates of analysis, and characterization data of all reference standards. Maintaining such traceability facilitates troubleshooting, such as discrepant method comparisons or shifts in assay performance with new reagent lots.

The role of the medical laboratory director

There are many guidelines that have been written by domain experts outlining the components and criteria for method validation and verification. However, there are times when the medical laboratory director has to make a clinical assessment on the performance of an assay. There are instances when collecting the recommended number of specimens for performing a method correlation might not be feasible. Examples include precious specimen types like various types of fluids or matrices (e.g., cerebrospinal fluid), and specimens positive for rare bacteria or viruses. The medical laboratory director would have to deem whether the number of specimens tested is acceptable.

In addition, most of the method validation criteria are applicable to quantitative and perhaps semiquantitative assays. Qualitative assays cannot be assessed for linearity nor can be assessed for precision using the CV. It is therefore the medical laboratory director's decision how to evaluate assay performance as well as which levels should be chosen for assessing the reproducibility of a positive or negative result. In addition, the cutoff concentration for either quantitative or qualitative assays should be carefully considered to meet clinical criteria. For example, many assays for prostate specific antigen have an LLOQ of $\sim$0.01 ng/mL, even though the medical decision point for biochemical recurrence is 0.2 ng/mL. Is it then appropriate to use 0.01 ng/mL as the assay cutoff or should results be reported as less than 0.2 ng/mL? Analytes such as cardiac troponin and human chorionic gonadotropin are reviewed frequently by emergency departments to make quick decisions for patient management. In these situations, a "negative" result is just as important as a quantitative level. It may then be important to evaluate a negative patient pool for precision and accuracy during the validation/verification process.

Collaborating with clinical experts and collecting feedback from clinicians are crucial to the method verification or validation process. Being familiar with the intended use of the assay and the medical decision points used by the clinical costumers contribute to the clinical validation of the method. Thus the medical laboratory director is not only responsible for assessing the analytical performance of the assay but also its performance within the appropriate clinical context.

References

[1] S.E.F. Melanson, N.I. Lindeman, P. Jarolim, Selecting automation for the clinical chemistry laboratory, Arch. Pathol. Lab. Med. 131 (7) (2007) 1063–1069.

[2] W. Jo, Basic Method Validation, third ed., Westgard Quality Corporation, Madison, WI, 1994.

[3] US Department of Health and Human Services, Medicare, Medicaid, and CLIA Programs: Regulations Implementing the Clinical Laboratory Improvement Amendments of 1988 (CLIA), Final. Rule Fed. Register 57 (1992) 7002–7186.

[4] R. Core Team. R: A language and environment for statistical computing. 2019; <https://www.R-project.org/>.

[5] R.J. McEnroe, P.A. Durham, M.D. Goldford, M.V. Kondratovich, S. Lababidi, R. Magari, et al., Evaluation of precision of quantitative measurement procedures; approved guideline, Clin. Lab. Stand. Inst. 34 (13) (2014).

[6] D.W. Tholen, M. Kroll, J.R. Astles, A.L. Caffo, T.M. Happe, J.L.F. Krouwer, Evaluation of the linearity of quantitative measurement procedures: a statistical approach; approved guideline, Clin. Lab. Stand. Inst. 23 (16) (2003).

[7] J.R. Budd, A.P. Durham, T.E. Gwise, B. Iriarte, A. Kallner, K. Linnet, et al. Measurement procedure comparison and bias estimation using patient samples; approved guideline. CLSI document EP09-A3. Clin. Lab. Stand. Inst. 2013. <https://clsi.org/media/2293/ep09ed3ce_sample.pdf>.

[8] J.S. Krouwer, G.S. Cembrowski, D.W. Tholen, Preliminary Evaluation of Quantitative Clinical Laboratory Measurement Procedures; Approved Guideline, Clin. Lab. Stand. Inst. 36 (34) (2014).
[9] K. Linnet, J.C. Boyd, Selection and analytical evaluation of methods - with statistical techniques, in: C.A. Burtis, E.R. Ashwood, D.E. Bruns (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, 5, Elsevier Saunders, St. Louis, MO, 2012, p. 7.
[10] P.C. Painter, Method evaluation and technology assessment, in: K. Lewandrowski (Ed.), Clinical Chemistry: Laboratory Management and Clinical Correlations, 1, Lippincott Williams and Wilkins, Philadelphia, PA, 2002, p. 122.
[11] S.P. Treon, L. Xu, G. Yang, Y. Zhou, X. Liu, Y. Cao, et al., MYD88 L265P somatic mutation in Waldenström's macroglobulinemia, N. Engl. J. Med. 367 (9) (2012) 826–833.
[12] FDA, CDER. Bioanalytical method validation guidance for industry biopharmaceutics bioanalytical method validation guidance for industry biopharmaceutics contains nonbinding recommendations, 2018. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/bioanalytical-method-validation-guidance-industry>.
[13] N.C. Hughes, E.Y.K. Wong, J. Fan, N. Bajaj, Determination of carryover and contamination for mass spectrometry-based chromatographic assays, AAPS J. 9 (3) (2007) E353–E360.
[14] J.S. Williams, S.H. Donahue, H. Gao, C.L. Brummel, Universal LC–MS method for minimized carryover in a discovery bioanalytical setting, Bioanalysis 4 (9) (2012) 1025–1037.
[15] D.A. Armbruster, D.B. Alexander, Sample to sample carryover: a source of analytical laboratory error and its relevance to integrated clinical chemistry/immunoassay systems, Clin. Chim. Acta 373 (1–2) (2006) 37–43.
[16] P.T. Vallano, S.B. Shugarts, E.J. Woolf, B.K. Matuszewski, Elimination of autosampler carryover in a bioanalytical HPLC-MS/MS method: a case study, J. Pharm. Biomed. Anal. 36 (5) (2005) 1073–1078.
[17] C.J. Mattocks, M.A. Morris, G. Matthijs, E. Swinnen, A. Corveleyn, E. Dequeker, et al., A standardized framework for the validation and verification of clinical molecular genetic tests, Eur. J. Hum. Genet. 18 (12) (2010) 1276–1288.
[18] D.H. Vu, R.A. Koster, A.M.A. Wessels, B. Greijdanus, J.W.C. Alffenaar, D.R.A. Uges, Troubleshooting carry-over of LC-MS/MS method for rifampicin, clarithromycin and metabolites in human plasma, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 917–918 (2013) 1–4.
[19] J.L. Hartley, A. Rashtchian, Dealing with contamination: enzymatic control of carryover contamination in PCR, PCR Methods Appl. 3 (2) (1993) S10–S14.
[20] T. Keller, T. Brinkmann, Proposed guidance for carryover studies, based on elementary equivalence testing techniques, Clin. Lab. 60 (7) (2014) 1153–1161.
[21] E. Gonzales, G. Ng, A. Pesce, C. West, R. West, C. Mikel, et al., Stability of pain-related medications, metabolites, and illicit substances in urine, Clin. Chim. Acta 416 (2013) 80–85.
[22] D.C. Mata, Stability of 26 sedative hypnotics in six toxicological matrices at different storage conditions, J. Anal. Toxicol. 40 (8) (2016) 663–668.
[23] J.F. Pierson-Perry, S.S. Altaie, S.J. Danielson, B.L. Jorgensen, B. Poetsch, R.M. Savol, et al., Evaluation of stability of in vitro diagnostic reagents; approved guideline, Clin. Lab. Stand. Inst. 29 (20) (2009) 40.
[24] M.P. Cornes, Exogenous sample contamination. Sources and interference, Clin. Biochem. 49 (18) (2016) 1340–1345.
[25] W. Guder, S. Narayanan, H. Wisser, B. Zawta, Endogenous and exogenous interferences, Samples: From the Patient to the Laboratory: The impact of Preanalytical Variables on the Quality of Laboratory Results, Wiley-VCH Verlag GmbH & Co, Weinheim, Germany, 2007.
[26] M.H. Kroll, R.J. Elin, Interference with clinical laboratory analyses, Clin. Chem. 40 (11 Pt 1) (1994) 1996–2005.
[27] U. Andreasson, A. Perret-Liaudet, L.J.C. van Waalwijk van Doorn, K. Blennow, D. Chiasserini, S. Engelborghs, et al., A practical guide to immunoassay method validation, Front. Neurol. 6 (2015) 179.
[28] L.F. Stevenson, S. Purushothama, Parallelism: considerations for the development, validation and implementation of PK and biomarker ligand-binding assays, Bioanalysis 6 (2) (2014) 185–198.
[29] B. DeSilva, W. Smith, R. Weiner, M. Kelley, J. Smolec, B. Lee, et al., Recommendations for the bioanalytical method validation of ligand-binding assays to support pharmacokinetic assessments of macromolecules, Pharm. Res. 20 (11) (2003) 1885–1900.
[30] B.D. Plikaytis, P.F. Holder, L.B. Pais, S.E. Maslanka, L.L. Gheesling, G.M. Carlone, Determination of parallelism and nonparallelism in bioassay dilution curves, J. Clin. Microbiol 32 (10) (1994) 2441–2447.
[31] M.-A. Valentin, S. Ma, A. Zhao, F. Legay, A. Avrameas, Validation of immunoassay for protein biomarkers: bioanalytical study plan implementation to support preclinical and clinical studies, J. Pharm. Biomed. Anal. 55 (5) (2011) 869–877.
[32] R. Cornelis, B. Heinzow, R.F. Herber, J.M. Christensen, O.M. Poulsen, E. Sabbioni, et al., Sample collection guidelines for trace elements in blood and urine. IUPAC Commission of Toxicology, J. Trace Elem. Med. Biol 10 (2) (1996) 103–127.
[33] C. McCudden, A.E. Axel, D. Slaets, T. Dejoie, P.L. Clemens, S. Frans, et al., Monitoring multiple myeloma patients treated with daratumumab: teasing out monoclonal antibody interference, Clin. Chem. Lab. Med 54 (6) (2016) 1095–1104.
[34] D. Lee, L.A.L. Bazydlo, G.M. Reisfield, B.A. Goldberger, Urine spiking in a pain medicine clinic: an attempt to simulate adherence, Pain. Med. 16 (7) (2015) 1449–1451.
[35] M.H. Kroll, M. Ruddel, D.W. Blank, R.J. Elin, A model for assessing interference, Clin. Chem. 33 (7) (1987) 1121–1123.
[36] G. Dimeski, Interference testing, Clin. Biochem. Rev. 29 (Suppl. 1) (2008) S43–S48.
[37] J.A. Dickerson, T.J. Laha, M.B. Pagano, B.R. O'Donnell, A.N. Hoofnagle, Improved detection of opioid use in chronic pain patients through monitoring of opioid glucuronides in urine, J. Anal. Toxicol. 36 (8) (2012) 541–547.
[38] R.J. McEnroe, G. Dimeski, P.A. Durham, J.J. Miller, W.G. Miller, V. Petrides, et al., Interference testing in clinical chemistry, Clin. Lab. Stand. Inst. EP07 (3) (2018) 109.
[39] P. Panuwet, R.E. Hunter, P.E. D'Souza, X. Chen, S.A. Radford, J.R. Cohen, et al., Biological matrix effects in quantitative tandem mass spectrometry-based analytical methods: advancing biomonitoring, Crit. Rev. Anal. Chem. 46 (2) (2016) 93–105.
[40] H. Trufelli, P. Palma, G. Famiglini, A. Cappiello, An overview of matrix effects in liquid chromatography-mass spectrometry, Mass. Spectrom. Rev. 30 (3) (2011) 491–509.

[41] B.K. Matuszewski, M.L. Constanzer, C.M. Chavez-Eng, Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC-MS/MS, Anal. Chem. 75 (13) (2003) 3019−3030.

[42] R. Bonfiglio, R.C. King, T.V. Olah, K. Merkle, The effects of sample preparation methods on the variability of the electrospray ionization response for model drug compounds, Rapid Commun. Mass. Spectrom 13 (12) (1999) 1175−1185.

[43] W. Dewé, M. De Smet, N. Evrard, B. Culot, M. Lastelle, G. Saucez, et al., Partial within-animal calibration: a new calibration approach in lead optimisation high-throughput bioanalysis, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 854 (1−2) (2007) 183−191.

[44] R. Dams, M.A. Huestis, W.E. Lambert, C.M. Murphy, Matrix effect in bio-analysis of illicit drugs with LC-MS/MS: influence of ionization type, sample preparation, and biofluid, J. Am. Soc. Mass. Spectrom. 14 (11) (2003) 1290−1294.

[45] D.A. Armbruster, T. Pry, Limit of blank, limit of detection and limit of quantitation, Clin. biochemist Rev. 29 (Suppl. 1) (2008) S49−S52.

[46] G.L. Long, J.D. Winefordner, Limit of detection a closer look at the IUPAC definition, Anal. Chem. 55 (07) (1983) 712A−724A.

[47] D.A. Armbruster, M.D. Tillman, L.M. Hubbs, Limit of detection (LQD)/limit of quantitation (LOQ): comparison of the empirical and the statistical methods exemplified with GC-MS assays of abused drugs, Clin. Chem. 40 (7 Pt 1) (1994) 1233−1238.

[48] Leito I., Kruve A., Rebane R., Oldekop M.-L., Evard H., Herodes K., et al. LC-MS Method Validation: Estimation of LOD [Internet]. University of Tartu. <https://sisu.ut.ee/lcms_method_-validation/93-estimating-lod>.

[49] Validation of Analytical Procedures: Text and Methodology: Q2 (R1). 2005. < https://www.fda.gov/regulatory-information/search-fda-guidance-documents/q2-r1-validation-analytical-procedures-text-and-methodology>.

[50] J. Tate, G. Ward, Interferences in immunoassay, Clin. Biochem. Rev. 25 (2) (2004) 105−120.

[51] D. Sartori, T. Lewis, A. Breaud, W. Clarke, The development of a high-performance liquid chromatography−tandem mass spectrometric method for simultaneous quantification of morphine, morphine-3-β-glucuronide, morphine-6-β-glucuronide, hydromorphone, and normorphine in serum, Clin. Biochem. 48 (18) (2015) 1283−1290.

Self-assessment questions

1. The within-run precision of your new assay is 15%. Your analytical goal for total imprecision is 15%. Which of the following is the best next step?
   - **a.** Assess the between-run precision over several days.
   - **b.** Perform the method comparison studies.
   - **c.** Perform the linearity or reportable ranges studies.
   - **d.** Review the data with the manufacturer.
2. The laboratory has been asked to report quantitative results for Hepatitis B surface antibody. Which of the following is NOT an acceptable source for linearity material?
   - **a.** quality control material
   - **b.** commercial linearity material
   - **c.** high patient pools
   - **d.** analyte-spiked samples
3. Your method comparison was done with 50 specimens spanning the medical decision points. Least-squares linear regression gives a slope of 1.20, an intercept of 0.1, and a correlation coefficient of 0.99. Which of the following best describes this analysis?
   - **a.** Additional specimens throughout the linear range are necessary.
   - **b.** The correlation coefficient indicates that the new method should be accepted.
   - **c.** A proportional bias of 20% between the methods exists.
   - **d.** The intercept suggests a significant constant bias between the methods.
4. U.S. federal law requires clinical laboratories to adhere to which of the following guidelines?
   - **a.** The Joint Commission (TJC)
   - **b.** College of American Pathologists (CAP)
   - **c.** Clinical Laboratory Improvement Amendment of 1988 (CLIA '88)
   - **d.** Clinical and Laboratory Standards Institute (CLSI)
   - **e.** Both c and d
5. Qualitative assays must follow the same validation parameters as quantitative assays.
   - **a.** True
   - **b.** False
6. If the laboratory changes the clinically reportable range (CRR) of an FDA-approved assay, it is now considered to be an FDA-modified test and must be validated using LDT parameters.
   - **a.** True, if the new CRR is outside of the manufacturer's defined AMR.
   - **b.** True, if the new CRR is within the manufacturer's defined AMR.
   - **c.** True, if the new CRR is either within or outside the manufacturer's defined AMR.
7. All of the following are examples of potential endogenous interferants EXCEPT _____.
   - **a.** human antimouse antibody
   - **b.** potassium
   - **c.** daratumumab
   - **d.** billirubin
   - **e.** albumin
8. Acceptable matrix effects in a lab developed test are ultimately determined by _____.
   - **a.** CLIA guidelines
   - **b.** CLSI guidelines
   - **c.** Medical Director discretion
9. Which of the following factors can influence analyte stability?
   - **a.** temperature
   - **b.** buffer composition
   - **c.** light exposure
   - **d.** time
   - **e.** all of the above

Answers

1. d
2. a
3. c
4. c
5. b
6. a
7. c
8. c
9. e

Chapter 5

Quality control

W. Greg Miller[1] and James H. Nichols[2]

[1]Department of Pathology, Virginia Commonwealth University; Richmond, VA, United States, [2]Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center; Nashville, TN, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Define the purpose of quality control for laboratory testing.
- Describe the principles of statistical process control.
- Identify the attributes and limitations of quality control materials.
- Explain the role of risk management and individualized quality control plans (IQCPs) for defining potential hazards and reducing errors in the laboratory testing process.
- Discuss the utility of proficiency testing.

Introduction

The total error of a laboratory test must be small enough to reflect changes in the biological condition in order to have value for clinical decision-making. The total error of a result is influenced by a number of factors including biological (physiological) variability within an individual; preanalytical variability in sample collection, transportation, processing, and storage; analytical variability in test performance; and interfering substances such as drugs or metabolic components.

This chapter addresses quality control (QC) of the analytical measurement process to ensure analytical variability meets accuracy and precision requirements that have been established for a measurement procedure and are considered appropriate for patient care. QC of the analytical measurement process is a component of an integrated quality management, or quality assurance, system, as shown in Fig. 5.1. The quality management system integrates good laboratory practices to assure correct results for patient care. Written standard operating procedures (SOPs) should be available to describe test methods that meet the organization's requirements for patient care, and SOPs should be followed by well-trained and competent personnel. Samples must be analyzed on well-maintained equipment that has been verified to meet manufacturer performance specifications, using validated reagents and QC samples that have been stored appropriately before and during use. Process control involves quality management of preanalytical, analytical, and postanalytical processes with documentation and review of results. Additional aspects of a quality management system should include information management (document and results control and record retention), occurrence management (incident response and resolution), assessment of quality benchmarks and process improvement, determination of customer service and satisfaction, and facilities for test performance in a safe manner.

QC is a part of the process control portion of the quality management system. An SOP should include all aspects of the analysis including the selection of QC materials and other monitoring parameters, how frequently to sample the measurement process, how to determine statistical parameters to describe a measuring system's performance, the QC plan and the acceptance criteria for QC results, corrective action when problems are identified, and the documentation and review processes. The SOP should also include who is authorized to establish acceptance criteria for the QC plan and the acceptance of results, who should review performance parameters including statistical QC results, and who can authorize exceptions to or modify an established QC policy or procedure.

The essential components of statistical process control are also shown in Fig. 5.1. The measurement system should be periodically evaluated by sampling its ability to produce the expected results for QC samples. The expected results are predefined and reflect the measurement variability when the system is operating correctly. If the QC results indicate a stable measurement process, then the patient results have a high probability of being correct. If the QC results fail evaluation criteria, then the patient results may not be reliable for clinical use. In the latter case, corrective action must be taken to fix the

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00005-3

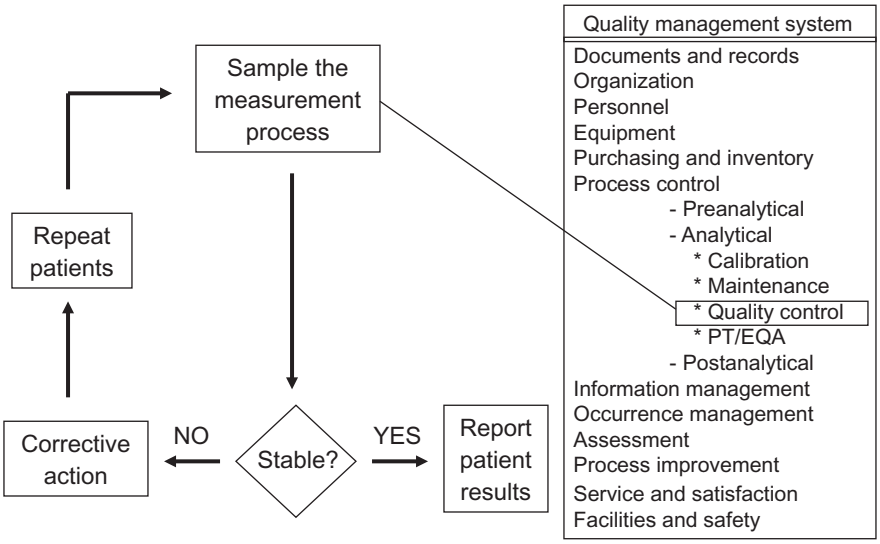


FIGURE 5.1 Overview of statistical process control (quality control) and its integration into a quality management system.

analytical process and patient samples must be reanalyzed for the interval since the last acceptable QC results. Good laboratory practice requires verification that a measuring system is performing correctly anytime patient samples are analyzed and results released to clinicians for management decisions.

Implementing statistical process control

Traditional statistical process control uses stable QC materials as surrogates for patient samples in a measurement process. Reliance on periodic analysis of QC samples does not require a comprehensive knowledge or assessment of potential hazards in the testing process, since this approach evaluates the aggregate effect of all components of the measurement system on the test result. Evaluation of the performance of QC samples provides a good means of detecting systematic errors that persist over time but may not detect random, sporadic errors in test performance. A complementary approach, discussed in a later section, assesses the potential hazards or weak points in each step of the testing process. Frequently occurring hazardous conditions or those with sufficient potential for patient harm require specific controls (that may be in addition to periodic measurement of QC samples) or actions to mitigate risk to the patient. Some devices have built-in checks or controls that monitor portions of the analytic process that may address specific potential hazards. Some of these internal built-in checks are performed with each test and may provide a better means of detecting random errors than reliance only on surrogate QC sample results. The compilation of all of the risk mitigation strategies, both built-in and external control processes, constitutes a QC plan for a specific measurement device that can reduce the residual risk of harm to a clinically acceptable level and ensure quality of test results.

Selection of quality control materials

Generally, two different concentrations of QC materials are necessary for adequate process control and are the minimum required by the Clinical Laboratory Improvement Amendments of 1988 (CLIA) regulations in the United States. For quantitative methods, QC materials should be selected to provide analyte concentrations that adequately monitor the analytical measurement range (also called the measuring interval) of the method. Most quantitative assays have a linear response over the analytical measurement range, and one can be confident that the performance is acceptable if the results near the lower and upper assay limits are acceptable. In the case of nonlinear method response, it may be necessary to use additional controls at intermediate concentrations. Critical concentrations for clinical decisions may also warrant QC monitoring. In the case of analytes that have poor precision at low to normal concentrations, such as bilirubin, the concentration should be chosen to provide an adequate standard deviation (SD) for practical evaluation. For procedures with extraction or pretreatment, at least one QC material should be employed to detect errors in the extraction or pretreatment step.

This chapter is primarily focused on QC procedures for quantitative methods. However, the principles can be adapted to most qualitative procedures. For qualitative tests, for example, urine drugs of abuse testing, both a negative and positive control are necessary, and controls with concentrations near the threshold are recommended to adequately control the method's ability to discriminate between negative and positive. For tests based on

qualitative interpretation of quantitative measurements, for example, drugs of abuse where a concentration is converted into a positive or negative result, it is also necessary to monitor near the threshold/cutoff concentrations to ensure appropriate discrimination between negative and positive responses. For qualitative procedures with graded responses, for example, dipstick urinalysis, a negative, and at least one positive with a value in the intermediate graded response region are appropriate. For qualitative tests based on other properties, for example, electrophoretic procedures and immunofluorescence, the QC procedure should appropriately verify the method's ability to discriminate normal from pathologic conditions.

The QC materials selected must be manufactured to provide a stable product that can be used for an extended time period, preferably 1 or more years. Use of a single lot for an extended period allows reliable interpretive criteria to be established that will permit efficient detection of an assay problem, avoid false alerts due to poorly defined acceptance ranges, and minimize the influence of reagent lot changes on the QC matrix and need for adjustment of target values.

Limitations of quality control materials

There are limitations inherent in currently available QC materials. One limitation is that QC materials are not always "commutable" with native clinical samples. A commutable QC material (or other reference material) is one that reacts in a measuring system to give a result that would be comparable to that expected for a native patient sample containing the same amount of analyte. Commercially available QC materials are typically noncommutable with native patient samples because the serum or other biological fluid matrix has been altered from that of a native patient sample [1]. The matrix alteration is due to processing of the biological fluid during product manufacturing, use of partially purified human and nonhuman analyte additives to achieve the desired concentrations, and various stabilization processes that can alter proteins, cells, and other components. The impact of matrix alteration on the recovery of an analyte in a measuring system is not predictable, is different for different lots of QC material, is different for different analytical methods, and can be different for different lots of reagent within the same analytical method [2].

A second limitation of QC materials is possible deterioration of the analyte during storage. Analyte stability during unopened refrigerated or frozen storage is generally excellent, but slow deterioration eventually limits the shelf life of a product and can introduce a gradual drift in control results. Analyte stability after reconstitution, thawing, vial opening, or exposure to light or air can be important sources of variability in QC results, and stability can vary substantially among analytes in the same vial. Repeated freeze/thaw cycles should be avoided. Users should limit the time spent at room temperature and the time spent uncapped with the potential for evaporation. An expiration time after opening should be established for each QC material and may be different for different analytes in the same control product. For QC materials reconstituted by adding a diluent, the vial-to-vial variability can be minimized by standardizing the pipetting procedure, for example, use the same pipette or filling device (preferably an automated device) and have the same person prepare the controls whenever practical.

Analyte concentrations in multiconstituent control materials may not be at optimal concentration levels for all assays. This limitation is caused by solubility considerations and potential interactions between different constituents, particularly at higher concentrations. Supplementary QC materials may be needed to adequately monitor the entire analytical measurement range when using a multiconstituent QC material to control several analytes at the same time.

Frequency to assay quality control samples

The frequency to assay QC samples is a function of several parameters:

- the analytical stability of the method,
- how much error can be tolerated without impacting patient care,
- the number of results produced in a period of time, and
- the need to verify and document the reliability of clinical results at the time they are reported.

The stability of the measurement system is the fundamental determinant of how frequently a QC sample needs to be assayed. The more stable the system, the less frequently a statistical process control needs to be performed. Minimum laboratory practice, consistent with CLIA Regulations Section 493.1256 [3], is to assay controls at least once per 24 hours, or more frequently if specified by the method manufacturer or if the laboratory determines that more frequent QC analysis is necessary to maintain the performance characteristics of a method. Some tests have more stringent requirements. For example, CLIA Regulations Section 493.1267 requires that blood gas measurements analyze at least one control every 8 hours that includes both high and low levels in the course of 24 hours; in addition, a control must be run with each patient sample unless the instrument automatically calibrates at least every 30 minutes. Methods that have automated built-in control procedures may use less frequent analysis of external QC materials (discussed in a later section).

The need to verify the clinical acceptability of results may support more frequent QC sampling than that based

TABLE 5.1 Common sources of measurement variability.

Source	Time interval for fluctuation	Likely statistical distribution
Pipette volume	Short	Gaussian
Instrument temperature control	Short or long	Gaussian or other
Electronic noise in the measuring system	Short	Gaussian
Calibration cycles	Short to long	Gaussian or shift (periodic step)
Reagent deterioration in storage	Long	Drift
Reagent deterioration after opening	Intermediate	Cyclic, periodic drift/step
Calibrator deterioration in storage	Long	Drift
Calibrator deterioration after opening	Intermediate	Cyclic, periodic drift/step
Control material deterioration in storage	Long	Drift
Control material deterioration after opening	Intermediate	Cyclic, periodic drift/step
Environmental temperature and humidity	Variable	Variable
Reagent lot changes	Intermediate to long	Shift (random step)
Calibrator lot changes	Intermediate to long	Shift (random step)
Instrument maintenance	Variable	Gaussian, cyclic, or shift (periodic step)
Deterioration of instrument components	Variable	Variable

strictly on method stability characteristics or on regulatory minimum requirements. More frequent QC sampling is appropriate with automated systems that release results continuously to minimize the probability to release incorrect results that could initiate inappropriate clinical action before an instrument problem is detected. For example, QC scheduled on a 24-hour cycle might be performed at 09:00 hours. If the QC results indicate a method problem, the erroneous condition could have started at any time during the previous 24 hours. If the problem had occurred at 15:00 hours the previous day, out-of-control results would have been reported for 18 hours, and the risk of harm to a patient from decisions based on those erroneous results must be considered. The cost of a medical error, or simply the cost of repeating the questionable patient samples, could be more expensive than more frequent QC analysis. Analytes, such as whole blood glucose and serum carbon dioxide, are not stable over 24 hours, and more frequent QC may be necessary to ensure sample viability should a specimen need to be repeated.

Establishing quality control target values and acceptance ranges that represent a stable measurement operating condition

QC target values and acceptable performance ranges are established to optimize the probability of detecting a measurement error that is large enough to impact clinical care. QC acceptable ranges should be sensitive enough to detect a problem while minimizing the frequency of false alerts. Prior to establishing QC values and the statistical parameters used to evaluate QC results, the measuring system must be correctly calibrated and operating within acceptable performance specifications. Some sources of measurement variability are listed in Table 5.1. Measurement variability can be affected by factors with short time-interval frequencies, many of which can be described by Gaussian error distributions, as well as intermediate and long time-interval factors that can cause cyclic fluctuations over several days or weeks, gradual drift over weeks or months, and more abrupt shifts in results. Acceptance criteria for QC results must adequately account for all sources of expected variability in results that are likely to occur when the measuring system is performing to specifications.

A QC material must have a reliable target value for an analyte. This requires adequate statistical sampling of both the QC material's vial-to-vial and open-vial stability, as well as the measurement variability during a period of time when the measuring system is calibrated correctly and is exhibiting the expected imprecision associated with a stable measurement condition. The objective to establishing a reliable QC target value is to include all sources of variability in the measurement process that will ensure a representative mean concentration. This objective is rarely met due to long-term variability components and the practical inability to account for all influences at the time of QC target value assignment. A Clinical and

Laboratory Standards Institute (CLSI) guideline for QC recommends the mean value from assaying a QC material on a minimum of 10 different days for an initial target value [4]. Note that it is preferable to use a single bottle of QC material for the expected duration of its open-vial stability in order to account for analyte stability effects rather than a fresh bottle on each day of testing. If a 10-day protocol is not practical, provisional, or temporary, target values can be established with less data but should be updated when additional replicate results are available. When applicable, more than one assay calibration event should be represented in the 10-day period to include the variability associated with the calibration process. The target value should be reassessed after additional QC data are available that include more sources of variability than can be included in 10 days.

Some QC materials are provided with manufacturer-assigned target values and ranges that can be used to confirm that a method meets the manufacturer's specifications. Such assigned values can be used initially by the laboratory, but both the target value and ranges should be reevaluated and adjusted by the laboratory after adequate replicate results have been obtained. Laboratory-defined target values and SDs ensure that the acceptance criteria used by a laboratory adequately reflect performance for the method in that laboratory. The acceptability limits suggested by a manufacturer may reflect sources of variability, such as between instruments and between reagent and calibrator lots, or may represent best-case or worst-case performance that may not be typical for that QC material in an individual laboratory. QC materials with assigned target values are also available from third-party manufacturers (i.e., manufacturers not affiliated with the method manufacturer). Caution should be used when employing assigned values for third-party QC materials, since the target values and ranges may have been established using protocols that do not adequately reflect specific measuring system performance. Noncommutability may further limit applicability of third-party QC values for some methods.

Once a target value has been assigned to a QC material, a range or SD must be established that represents the typical imprecision of the measuring system when it is performing according to its manufacturer's specifications. Although there are non-Gaussian components of variability in QC results, an SD is the conventional way to express method variability because the statistical packages in instrument and laboratory computer systems are designed for QC data analysis based on mean and SD parameters. An SD based on the data from a 10-day target value assignment, or a 30-day monthly summary, has a large uncertainty and cannot reflect all sources of variability that need to be accommodated in the QC result evaluation process [4]. When previous experience with a method exists, the SD or range of QC results for stable measurement performance can be established from the cumulative SD over a 6- to 12-month period, which will include most sources of expected variation. Note that the cumulative SD is not established from the average of monthly QC statistical summaries. Different sources of long-term variability, which occur at different times during the use of a measuring system, may not be represented in a monthly SD for QC results. However, the cumulative SD includes contributions from most sources of variability as they occur and as they are reflected in the individual QC results. Consequently, the cumulative SD will typically be larger than the monthly values and better represent the actual variability of the measuring system. If the imprecision expected during stable operation is underestimated, then the acceptable range for QC results will be too narrow and the false-alert rate will be unacceptably high. If the imprecision is overestimated, then the acceptable range for QC results will be too large and the QC process will be less sensitive to detect a measurement error.

It is important to include all valid QC results in the calculation of a QC range. A valid QC result is one that was, or would have been, used to verify acceptable measuring system performance and reporting of patient results. Only QC results that would have been responsible for holding the release of patient results should be deleted (with documentation) from summary calculations. Selective editing of QC results underestimates the true method of SD, so QC ranges will be too narrow and will generate increased false QC alerts with concomitant reduction in the effectiveness of the statistical process control.

When a method has been established in a laboratory, and a new lot of QC material is being introduced, the target value for the new lot should be used with the cumulative SD from the previous lot to establish an acceptable range for the new lot of QC material. This practice is appropriate because the measurement imprecision is a characteristic of the measuring system. The measuring system imprecision is unlikely to change with a different lot of QC material. If the target values for the old and new lots are substantially different, there may be a different imprecision observed, and an adjustment to the QC SD may be necessary as additional experience with the new lot of QC material is accumulated.

When a new method is introduced for which there is no historical performance information, then the SD for stable performance must be established using data from the method validation and target value assignment of the QC materials. Additional information on expected method performance can be obtained from the method manufacturer, the package insert, and the method's performance in interlaboratory QC and/or proficiency testing (PT)

programs. For a new method, the initial SD and QC range evaluation criteria will need to be monitored closely, and adjustments made after more data have been collected to include intermediate and long-term sources of method variability.

Establishing acceptance criteria to evaluate quality control results

Operating conditions must be established for each measuring system before suitable QC interpretive criteria or rules can be selected. Establishing acceptance criteria to evaluate QC results is a balance between the sensitivity of detecting a significant analytical error condition and the frequency of false alerts. Various statistical models and assumptions regarding distribution of errors have been used to establish interpretive rules for QC results [5–7]. The probability for any interpretive rule to detect a measurement error increases as the magnitude of the error increases. When evaluating interpretive rules, it is common practice to express the magnitude of an error in multiples of the SD for the method (i.e., SD intervals, SDIs, or z-score).

A conventional way to express QC interpretive rules is with an abbreviation nomenclature summarized in Table 5.2. Note that fractional SDIs are permissible, as in the $2_{2.5S}$ rule. For example, consider a 1_{2S} interpretive rule that would be violated if a single QC result was >2 SD from the target value. A power function graph [7] shows that a 1_{2S} rule has a 35% probability to detect a systematic error that is 1 SD in magnitude and a 78% or 98% probability to detect systematic errors that are 2 or 3 SD in magnitude, respectively. However, it is not recommended to use a 1_{2S} rule because, due to the inherent imprecision, a 1_{2S} rule has an almost 10% probability of falsely indicating an error condition when there is zero systematic bias. Trend detection techniques such as cumulative sum or exponentially weighted moving averages are more powerful than rules based on the number of results that exceed a particular SDI and are recommended when computer systems can support their use [5,8].

A combination of interpretive rules applied simultaneously to a set of QC results will optimally detect errors with the lowest false-alert rates [6,7,9]. For example, a $1_{3S}/2_{2S}/R_{4S}$ multirule applied to two QC results identifies an error condition if one control exceeds ±3 SD from the target value, if two controls exceed ±2 SD in the same direction from the target value, or if the range between two controls exceeds 4 SD. In this multirule example, the 1_{3S} component is sensitive to large systematic bias, the 2_{2S} component is sensitive to systematic bias, and the R_{4S} component is sensitive to imprecision. This multirule has a low false-alert rate and a better probability to detect an error condition than the individual components alone. A number of multirule combinations can be used as appropriate for the performance characteristics and clinical requirements of a measuring system.

A challenge when selecting QC interpretive rules is that all sources of variability listed in Table 5.1 occur in most contemporary automated measuring systems and introduce long-term cyclic and step fluctuations in performance. These sources of variability are not adequately described by Gaussian-based QC interpretive rules. Since almost all instrument and laboratory computer system QC software is based on SDI (z-score) interpretation, it is necessary to estimate an SD that approximates the long-term variability from Gaussian and non-Gaussian contributions. Thus data collected over a significant time period are necessary to ensure the SD and QC ranges represent typical

TABLE 5.2 Abbreviation nomenclature to express quality control (QC) rules assuming Gaussian (normal) distribution of imprecision.

Rule	Meaning	Detects
1_{2S}	One observation exceeds 2 SD from the target value	Imprecision or systematic bias
1_{3S}	One observation exceeds 3 SD from the target value	Imprecision or systematic bias
2_{2s}	Two sequential observations, or observations for two QC samples in the same run, exceed 2 SD from the target value in the same direction	Systematic bias
$2_{2.5s}$	Two sequential observations, or observations for two QC samples in the same run, exceed 2.5 SD from the target value in the same direction	Systematic bias
R_{4S}	Range between two observations in the same run exceeds 4 SD	Imprecision
10_m	Ten sequential observations are on the same side of the target value (mean)	Systematic bias and trend (not recommended)
8_{1S}	Eight sequential observations for the same material exceed 1 SD in the same direction from the target value	Systematic bias and trend

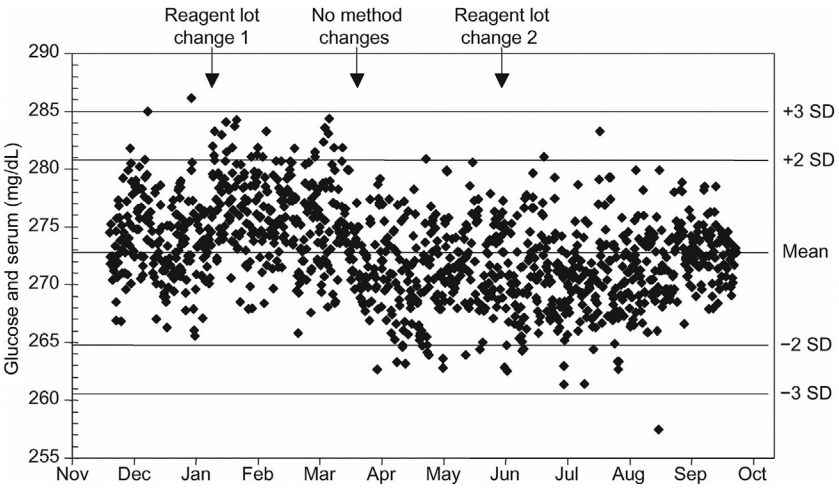


FIGURE 5.2 Levey–Jennings plot of quality control results ($n = 1232$) for a single lot of quality control material used for a 10-month period. The mean and standard deviation are cumulative values from all data.

variability when the method is stable and performing correctly.

Fig. 5.2 shows results for a single lot of QC material over a 10-month period for an automated glucose measuring system. The glucose method stability and performance over the 10 months were considered acceptable for clinical use. If acceptance criteria were based on the SD from the initial 30-day period, the subsequent long-term QC result fluctuations would have caused a large number of false alerts with significant effort and expense for troubleshooting what was actually inherent variability during stable method performance. The SD lines shown in Fig. 5.2 were determined from the entire 10 months of data. Examination of these data shows several fluctuations that cannot be described by a Gaussian statistical model. The first reagent lot change caused a step shift to higher values, but the shift was small enough to not require a target value adjustment (see later section). The second reagent lot change had no effect on the QC results. Between March and April, there was a transition to lower results that did not correspond to any reagent lot change, maintenance, or calibration events. Throughout the 10-month period, there are intervals of several weeks duration when the imprecision was either better or worse than other time periods emphasizing the importance to use a cumulative SD over a sufficient time interval that includes most sources of variability.

An empirical approach is frequently necessary to establish QC interpretive rules that will adequately encompass the observed variability in a measuring system. An example of an empirically developed multirule based on the data in Fig. 5.2 is $1_{3S}/2_{2.5S}/R_{4S}/8_{1.5S}$, in which 1_{3S} detects imprecision or bias, $2_{2.5S}$ detects systematic bias, R_{4S} detects imprecision, and $8_{1.5S}$ detects a trend in bias based on eight consecutive results that exceed 1.5 SD. This multirule was selected based on the clinical requirements for patient care, observed long-term measuring system performance, and the need to identify potential measurement issues with a false-alert rate $<1\%$. This multirule had 0.6% false alerts when applied to the data in Fig. 5.2 with a target value assigned as the mean value from the November-to-January period and the SD for the 10-month period to represent overall imprecision. If a 2_{2S} rule had been used instead of a $2_{2.5S}$ rule (sequential observations in this example), the false-alert rate would have increased from 0.6% to 1.8%. An $8_{1.5S}$ rule was selected to provide detection of bias trends because it had zero false alerts compared to 0.5% false alerts for an 8_{1S} rule. The $8_{1.5S}$ rule value was small enough that a bias trend of this magnitude would not adversely affect clinical interpretation of patient results. A rule that evaluates sequential results on one side of the mean can have a high false-alert rate when a measuring system's actual error distributions are not perfectly Gaussian. For example, a 10_m rule for the data in Fig. 5.2 would have increased the false-alert rate by 10.6%, which is why a 10_m rule is not recommended. Many contemporary analyzers are very stable and may produce QC results on one side of the mean for extended periods of time; however, if the magnitude of the difference from the target value is small, it may not indicate a problem with clinical interpretation of the patients' results.

In practice, empirical judgment is frequently necessary to set QC rules that fit the actual data accumulated over a long enough time period to adequately describe the variability observed when a measuring system is working correctly. Caution should be used when selecting QC interpretive rules based only on Gaussian models of imprecision because the rules may not accommodate all types of variability observed for many measuring systems. Whatever statistical approach is used, the balance between false alerts and the probability of error detection is improved when multiple rules are used in combination. When establishing rules to interpret QC results, statistical

process control can verify that a measuring system is producing results within its expected variability only at a point in time compared to when the system was properly calibrated and in a stable operating condition. QC rules are intended to detect bias and imprecision that is large enough to require correction before patient results are reported. Random events, for example, a temporary clot in a pipette, that do not persist until the next QC sample is measured will not be detected by external surrogate QC sample evaluation.

A laboratory director may determine in the process of reviewing statistical parameters for QC data that a measuring system's variability is too large to meet medical requirements. If the measuring system is performing in a stable condition within its manufacturer's specifications, the observed variability is an inherent property of the current technology. In this case, the only solutions are to improve the measuring system performance or to use a different method. If the measuring system performance cannot be improved and a better method is not available, the laboratory must accept the limitations and communicate them to patient care providers. In this circumstance, the QC ranges or interpretive rules should not be made artificially stringent. This incorrect approach will not improve measuring system performance but will increase the QC false-alert rate and decrease the efficiency and practicality of the statistical QC process.

Corrective action when a quality control result indicates a measurement problem

A QC alert occurs when a QC result fails an evaluation rule, indicating that an analytical problem may exist. A QC alert should occur only when there is a high probability the measuring system is producing results that are unreliable for patient care. When this condition occurs, corrective action should be taken to investigate the cause for the QC alert. Fig. 5.3 presents a generalized troubleshooting sequence. When a QC alert occurs, the first action is to discontinue assaying patient samples. Repeating the QC measurement on the same sample is not recommended, because, with properly designed acceptance criteria, it is more likely that a measurement system problem exists than the QC result was a statistical outlier. However, QC materials can deteriorate after opening due to improper handling and storage or due to labile analytes. Thus repeating the measurement on a new vial of the QC material may establish that the alert was caused by a deteriorated QC material rather than a method problem. In this situation, if the QC alert is resolved, testing of patient samples can resume. One caution is to consider a developing trend. If the repeat result is very near the acceptability limits, it means the measuring system's results are close to unacceptability and may indicate an impending problem that should be investigated as soon as possible. It is important that the multirule sequence include a

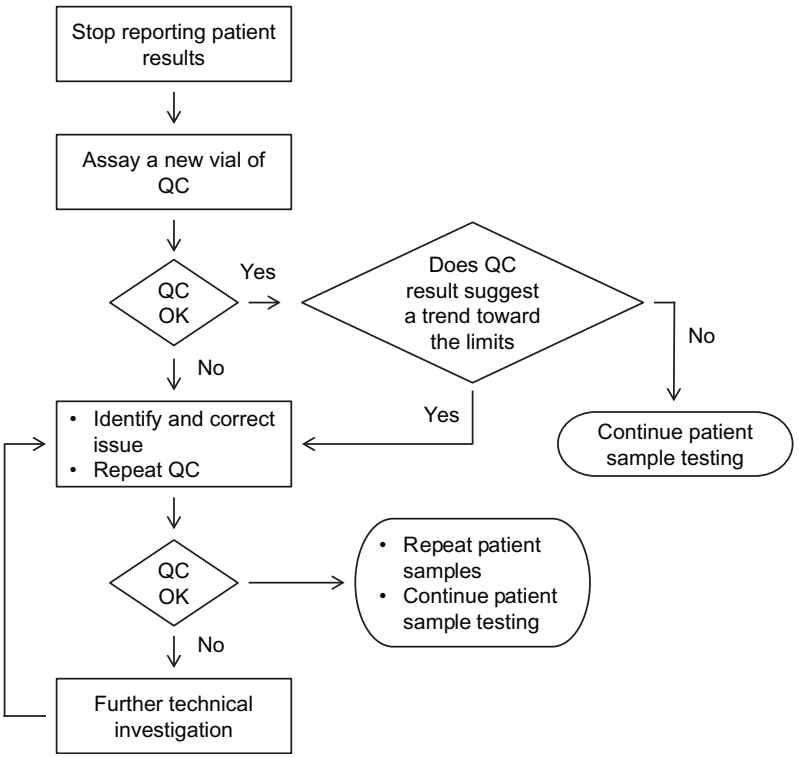


FIGURE 5.3 Generalized troubleshooting sequence following a quality control alert.

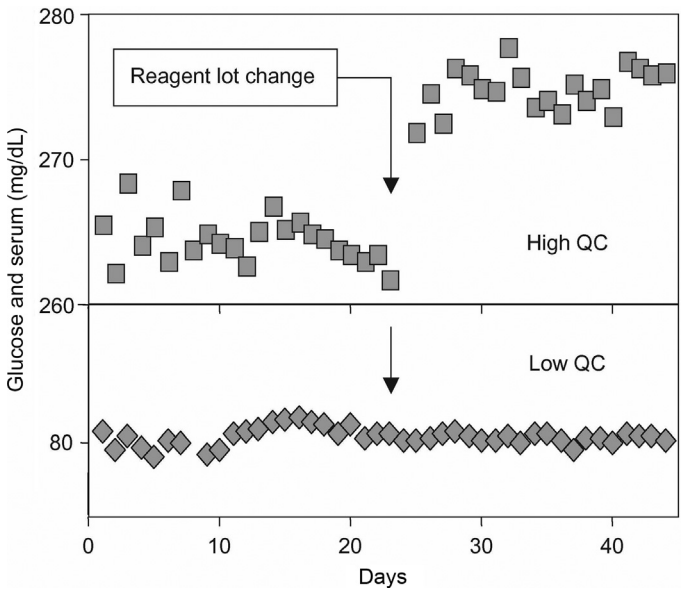


FIGURE 5.4 Levey–Jennings graph showing impact of a reagent lot change on matrix interaction with quality control materials.

trend-detecting QC rule, such as the $8_{1.5S}$ (in the previous example), an exponentially weighted moving average, or a cumulative sum.

When repeat testing of a new QC sample does not resolve the alert situation, the instrument and reagents should be inspected for component deterioration, mechanical problems, or other issues. When the problem is identified and corrected, controls should be measured to verify assay performance has acceptably returned to stable conditions. In many cases, the method will require recalibration before QC results can confirm correct method performance. After confirmation of correction, patient samples since the time of the last acceptable QC results should be repeated. The laboratory director should establish acceptable performance criteria for results from repeat testing of patient samples in order to determine whether the original test results can be reported or whether a corrected result should be reported. The criteria for acceptability of repeated tests are based on measuring system performance, the patient population, and the risk of harm to a patient.

In some cases, there may not be adequate volume left in patient samples (quantity not sufficient; QNS) for repeat testing. In these situations, no results can be reported unless the impact of the error on the original results was small enough to have an insignificant effect on clinical interpretation. One way to evaluate the clinical impact of a measuring system problem is to repeat those patient samples that have adequate volume. The repeated samples must represent the concentration range of the QNS samples, must represent the time span since the previous acceptable QC results, and should include a suitable number of samples originally assayed while the method was in the unacceptable condition to ensure representative assessment of the interval when there was a measurement defect. If the repeat results for this sample group are within established performance criteria for repeat testing of patient specimens, the original results for the QNS samples can be reported. Otherwise, the original results for the QNS samples are considered erroneous, and a corrected report should indicate no results and that the original results were incorrect.

Verifying quality control evaluation parameters following a reagent lot change

Changing reagent lots can have an unexpected impact on QC results. Careful cross-over evaluation of QC target values is necessary. Because the matrix interaction between a QC material and a reagent can change with a different reagent lot, the QC results may not be a reliable indicator of a method's performance for patient samples following a reagent lot change [2]. In the example in Fig. 5.4, the QC values for the high-concentration control shifted following the change to a new lot of reagents, but there was no change in results for the low-concentration control. A comparison of results for native patient samples assayed using the old and new reagent lots demonstrated equivalent patient results using either lot of reagents. Consequently, the change in QC values was due to a difference in matrix interaction between the high-concentration QC material and the two reagent lots.

When changing reagent lots, laboratories should be aware that the target value for a QC material may be affected. Target values for QC materials should be adjusted when a different matrix interaction is observed with a new reagent lot. First, patient sample results measured using the old and new reagent lots should be

compared to verify equivalent results are produced. If a problem is identified, the calibration of the new reagent lot should be investigated and corrected, or the possibility of defective reagents considered. The native patient sample results, not the QC results, are the basis to verify that results using the new reagent lot are consistent with those using the previous reagent lot. CLSI document EP26 provides guidance for verifying reagent lot changes and recommends a minimum of three patient samples or more depending on the number of critical clinical decision values and the imprecision of a measuring system [10]. EP26 also includes guidelines for acceptance criteria related to the magnitude of a difference that is clinically meaningful and the imprecision of a measuring system. CLSI document C24 also includes information on the influence of reagent lot and other changes on QC target values [4].

Once patient results are shown to be equivalent within clinically acceptable limits, the results from QC samples should be evaluated to determine whether the target values have changed with the new lot of reagent. If a QC target value has changed, that value should be adjusted to reflect the new value to be used with the new lot of reagents to keep the QC interpretive rules centered on the appropriate QC target value. This target value adjustment will allow the same QC interpretive rules to be utilized. In making this adjustment, the laboratory is compensating for an altered matrix interaction of the QC material with a different reagent lot. The SD used to evaluate QC results will not typically change when a new lot of reagent is put into service. The SD represents the expected variability when the method is stable and performing according to specifications. In most cases, the variability of a method will be the same with any lot of reagent(s); however, there may be occasional exceptions requiring adjustment of the SD, for example, if the reagent is reformulated and performance changes.

Verifying method performance following use of a new lot of calibrator

When implementing a new lot of calibrator, with no change in reagent, there should be no change in matrix interaction between the QC material and the reagent. In this case, the QC results provide a reliable indication of calibration status with the new lot of calibrator. If the QC results indicate a bias following use of a new lot of calibrator, the problem is caused by the new lot of calibrator and needs to be addressed to ensure consistent results for patient samples.

Some manufacturers package kits that include reagents, calibrators, and QC materials together. In this case, the QC results could fail to identify a calibration shift when a new kit is opened. Native patient samples should be analyzed with the old and new kits to verify consistency of patient results. When possible, QC materials that are independent of the kit lot are recommended. Laboratories should also avoid changing lots of QC material at the same time as changing lots of reagent or calibrators. The analysis of native patient samples, except for unstable analytes, provides a reliable approach to verify the consistency of results following changes in lots of reagents or calibrators.

Calibration issues in quality control

Calibration of the analytical measurement system is a key component to achieve quality results. The principal reason to perform statistical process control is to verify that the calibration and the variability of the analytical system remain within limits expected for stable measuring system performance. Specific techniques for calibration are unique to individual measuring systems and will not be covered here. However, some general principles for implementing calibration procedures, and for verifying calibration uniformity among multiple methods, can contribute to stability and clinical reliability of laboratory results.

Calibration is most often performed by the laboratory using calibrator materials provided by the measuring system manufacturer. In some cases, for example, point-of-care devices, measuring systems are calibrated during the manufacturing process, and the laboratory performs a verification of that calibration. In either situation, traceability of result accuracy to the highest order reference system is provided by the measuring system manufacturer. The measuring system manufacturer's calibrator material and assigned target values are designed to produce accurate results with native patient samples assayed using that particular manufacturer's measuring system. One manufacturer's product calibrator is not intended to be used with other measuring system, and laboratories should not use calibrator materials intended for one measuring system with any other measuring system because the matrix interaction with different reagents will produce a different response. Use of a calibrator with a measuring system for which it was not specifically intended can produce miscalibration and erroneous patient results.

National and international reference materials are available for some analytes. In many cases, these reference materials are intended for use with higher order reference measurement procedures and may not be suitable to use directly with routine clinical laboratory measuring systems. Laboratories should not use national or international reference materials to directly calibrate a clinical laboratory measuring system (or to verify the calibration of a clinical laboratory measuring system) unless

commutability with native patient samples has been verified for that specific reference material and clinical laboratory measuring system (see Chapter 17: Harmonization of results among laboratories, for additional information on calibration traceability and commutable reference materials). The reference material certificate of analysis should be reviewed for commutability documentation. If the reference material is noncommutable, the clinical laboratory measuring system could actually be miscalibrated and produce erroneous results for patient samples [11,12].

Optimal long-term stability of patient results is achieved if recalibration is performed only when the relationship between the analytical measuring system response and specimen concentration has changed. Performing a recalibration when the calibration relationship between measurement signal and concentration has not changed can be a source of imprecision in results. This imprecision occurs because the new calibration will produce a slightly different relationship between analytical system response and specimen concentration. CLIA Regulations Section 493.1255 requires calibration or calibration verification at least every 6 months or more frequently if recommended by the measuring system manufacturer [3]. Recalibration should be performed when necessitated by drift (which can be detected by shifts in QC results or other monitoring parameters), by changes in reagent lots, or after significant instrument maintenance. When there has been no change in measuring system performance, verification of the current calibration may be preferred rather than performing a recalibration.

One common procedure to verify calibration is to analyze the calibrator materials as "unknowns." Recovery of the target values indicates the measuring system calibration has not changed. There is no reason to perform a recalibration because the same relationship would be reestablished between measurement signal and the concentration of analyte in the calibrators. The laboratory must establish criteria for calibration verification (i.e., agreement between the assayed result and the calibrator target value). Conservative criteria for agreement, such as ± 1 SD from the target value, should be considered to avoid misinterpretation of the calibration status of a method. Calibration can also be verified by performing a method comparison using native patient samples measured with another measuring system known to be correctly calibrated.

Another approach to verify calibration is to analyze control materials that are specifically intended for this purpose. Some manufacturers produce control materials, called trueness controls, designed for calibration verification. Such manufacturer-provided materials typically have matrix characteristics and target values specifically designed for the measuring systems claimed in the package insert and cannot be used with any other measuring systems. These calibration verification materials may have target values certified by the manufacturer to be suitable for all reagent lots, or they may be specific for stated reagent lots. Third-party QC materials are usually not suitable for calibration verification. These materials are typically not validated for commutability with native clinical specimens for use with different measuring systems and do not have target values that are traceable to reference systems.

Development of a quality control plan based on risk management

Historically, the quality of laboratory results was ensured by periodic testing of QC samples to verify stability of measuring system performance. QC samples act as surrogates for patient samples and have target values that when recovered upon periodic analysis verify acceptable measuring system performance. Many regulatory agencies have required periodic analysis of QC samples as a means of ensuring quality. For instance, CLIA requires the analysis of at least two levels of controls each day of patient testing (more frequently for blood gas and coagulation devices). This QC strategy is sensitive to changes in the measuring system that persist in time (systematic errors). But, external QC samples do a poor job at detecting random errors that occur sporadically, at one point in time, or for a single patient sample (i.e., hemolysis, clots, and drug interferences).

Newer laboratory instrumentation utilizes a variety of control processes to ensure the quality of patient results. Some of these controls are built-in checks performed by the device to verify instrument function, such as electrical checks, optical checks, and biosensor checks. For example, the control line which develops on a lateral flow pregnancy or drug test evaluates kit storage viability, appropriate sample volume, sample application, and adequate timing. Other devices store on-board QC materials that are analyzed periodically by the measuring system to automate QC sample measurement. Still other devices require the operator to perform an action, like verifying the integrity of a temperature check (a simple device that changes color if frozen or heated during transit) when receiving shipments of reagents and calibrators. Some of these control processes are conducted with each analysis, providing a better means of detecting certain types of random errors than periodic analysis of external QC samples. Each control process is directed at minimizing the probability of certain errors, and devices have different risks for different types of errors. Thus each device may have different control processes to mitigate risk. Consequently, no single QC plan is totally adequate for all devices given

the complexity and variety of laboratory instrumentation that is available on the market today.

Use of risk management principles allows the development of a QC plan that is customized to an individual device, laboratory setting, and clinical application. CLSI guideline EP23 uses risk management to develop a laboratory QC plan [13]. EP23 is intended for clinical laboratories and provides guidance on using the manufacturer-provided information in the package insert, instructions for use and owner's manual, the risk of errors when testing with the device, and engineered control processes available on the instrumentation to develop custom, individualized QC plans (IQCPs).

Fig. 5.5 shows the process for developing a QC plan [13]. The laboratory starts by collecting information about the medical requirements of the test, local and regulatory QC requirements, the measuring system, and the heath care and laboratory settings. This information is processed using a risk assessment to develop a QC plan that will meet patient care needs and local QC regulations. Potential hazards or weaknesses in each step of analysis are identified. These hazards are assessed for frequency of occurrence and potential harm to the patient. Mitigation or control processes are required for those hazards that have greatest risk of harm to the patient or highest frequency of occurrence in order to reduce the residual risk of an error to a clinically acceptable level. The collection of all the control processes that are executed by the laboratory (surrogate QC sample testing plus other risk mitigation procedures) becomes the laboratory specific QC plan for that device or measuring system. The laboratory can never predict every possible hazard that could occur with a laboratory device, so it is never possible to entirely eliminate risk. There will always be some residual risk that the laboratory should manage through its control processes to minimize errors to a clinically acceptable level. Once implemented, the QC plan is monitored for the occurrence of unexpected errors or complaints. When errors or trends are identified, new control processes are defined to further mitigate and minimize risk, and the original QC plan is modified. In this manner, the laboratory responds to issues with corrective and preventive actions and cycles of continuous quality improvement.

The QC plan provides documentation of the reasoning for laboratory QC processes for a specific device. The QC plan defines what potential hazards the laboratory considered, the mitigations that will reduce risk of those hazards to a clinically acceptable level, the frequency of QC processes, and how the QC plan will be monitored (baseline measures for improvement). The QC plan provides a written outline of the laboratory's strategy for controlling the quality of a laboratory method and provides the rationale for control process improvement. Regulatory inspectors will find the QC plan sets laboratory policy for a measuring system's quality by defining how the laboratory will manage hazards, and inspectors can look to specific SOPs to determine whether the laboratory is following the elements of the QC plan.

Under the CLIA interpretive guidelines implemented on January 1, 2016, by the Centers for Medicare and

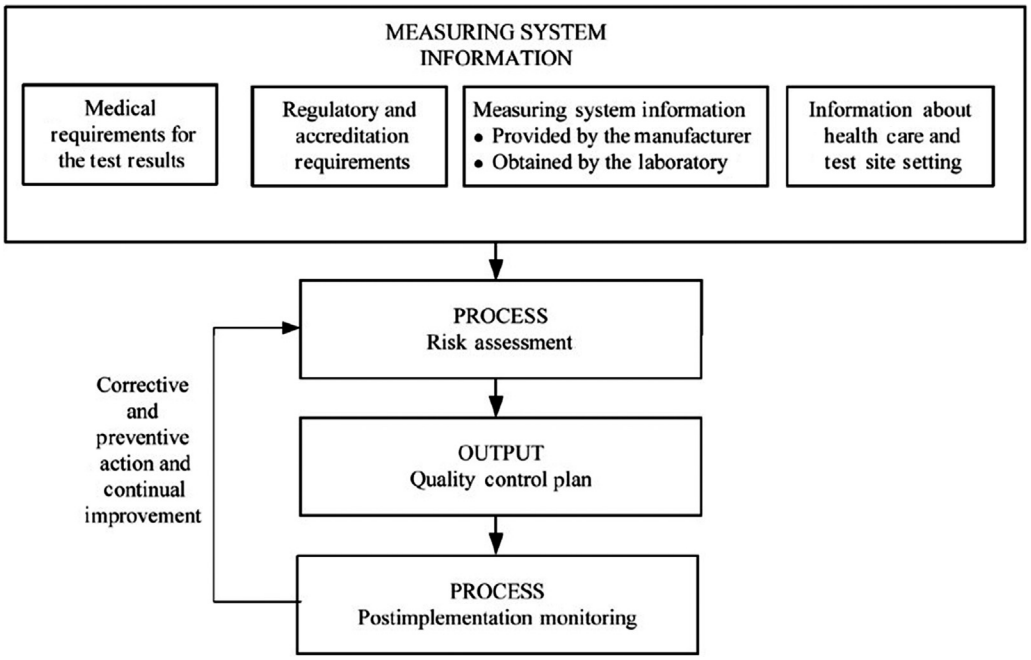


FIGURE 5.5 Process to develop and continually improve a quality control plan. *Reproduced from the EP-23-A Document with permission from the Clinical and Laboratory Standards Institute.*

Medicaid Services (CMS), laboratories in the United States have two options: (1) perform a minimum of two levels of QC for each test daily; or (2) develop an IQCP [14]. CMS is initially requiring an IQCP for CLIA nonwaived tests. IQCPs will allow laboratories to find the right balance of surrogate sample QC frequency with manufacturer built-in QC and control processes engineered into the device. Laboratories will be able to utilize the device control processes to reduce the frequency of surrogate sample QC for measuring systems that have minimal risk. In other situations, laboratories may find that they will need to increase the frequency of surrogate sample QC as well as implement additional control processes for high-risk measuring systems to achieve a clinically acceptable level of risk. Inspectors will be reviewing three items as part of the IQCP inspection process: (1) the laboratory's risk assessment; (2) the IQCP summary; and (3) the laboratory's quality assessment (e.g., the laboratory's quality benchmarks monitored to prove the effectiveness of the IQCP). As the philosophy of risk management is more widely adopted, laboratories can expect to see additional changes to regulations that will allow a variety of QC processes to be utilized in various combinations to manage measuring system quality and more reliance on laboratory specific IQCPs to define a laboratory's QC strategy.

Reviewing the quality control plan

A QC plan should be reviewed at least annually or whenever troubleshooting of QC errors or trends indicates the need to change the QC strategy. To develop a QC plan requires review of the many steps of the testing process looking for potential hazards or weaknesses where errors may occur. In a central laboratory with the analyzers all located in the same space, the analytical process may be more uniform than a point-of-care testing program with dozens of devices spread throughout an institution and hundreds of staff conducting testing. Discussion of the testing process with representative staff can uncover variations that may present the potential for errors in one location that may be less likely in other locations. Consider specimen labeling, for instance. Printing multiple specimen labels in advance before specimen collection may seem to be efficient but also presents the potential to mix-up labels between patient samples compared to on-demand printing of the label at the time of specimen collection. This difference in process requires an extra control step where staff must verify the correct specimen label matches the right patient and the right collection container with suitable preservative when needed. This verification step may be overlooked if printing occurs in advance. Printing labels on-demand during specimen collection has a much lower probability of specimen mix-up and the chance of human error during labeling.

Minor variations in preanalytical practices can occur throughout the testing process. Individual operator technique, specimen processing, aliquotting, dilutions, reagent preparation, and calibration all present a potential for error. QC specimens, if handled in the same manner as patient specimens, may detect some of these operational variations, especially if the technique leads to errors or increased variability in results. However, QC is often handled differently than patient samples and thus may not identify all sources of potential hazards. The management and handling of QC should therefore be considered in developing or revising a QC plan.

Review of the QC plan and discussion of the testing process with those involved is also an opportunity to uncover QC efficiencies that may not have been considered. Many point-of-care devices act as test readers and are simple voltmeters or light photometers with no moving parts. The chemistry of these devices is in the cartridges or test strips. QC can be performed on every device and a representative test cartridge every 24 hours, but with dozens of devices and locations to manage, this frequency may be excessive given the manufacturer control processes that are engineered into each test cartridge/strip. A urine pregnancy test, for example, has a control line that develops with each test. Analyzing a QC sample on these devices verifies the viability of that lot of test cartridges/strips during storage but can be duplicative to the internal control on that cartridge/strip. A QC plan may therefore consider utilizing the internal control on each test as the process for verifying chemical viability while employing the external QC samples as an event-driven process for confirming viability of new shipments, comparability of new lots (together with patient sample correlations), and detection of any deterioration during storage (periodic weekly or monthly analysis).

Testing surrogate sample QC materials cannot be billed to a patient or reimbursed by insurance but is required to manage the quality of the testing process. Large institutions that share a common lot of cartridges/strips might consider performing QC on a subset of analyzers, since the QC is detecting the chemistry of the test while the devices are internally verifying the voltage and photometer readouts when devices are turned on. Table 5.3 shows that an institution that has 20 readers receives quarterly shipments and conducts monthly QC on two lots of cartridges/strips could achieve significant efficiencies versus performing QC for every device and shipment. Such considerations are the heart of strategies for developing a robust QC plan that meets clinical needs but is also cost-effective and optimizes available resources while ensuring high-quality test results.

TABLE 5.3 Example comparing performing quality control (QC) for every point-of-care device versus a subset of devices[a].

QC for every device
20 readers × 4 quarterly shipments × 2 levels of QC = 160 QC tests
20 readers × 12 monthly QC × 2 levels of QC × 2 lots cartridges/strips = 960 QC tests
Total annual = 1120 QC tests
QC for a subset of devices
3 readers × 4 quarterly shipments × 2 levels of QC = 24 QC tests
3 readers × 12 monthly QC × 2 levels of QC × 2 lots cartridges/strips = 144 QC tests
Total annual = 168 QC tests
Savings of 952 QC tests or 85% reduction QC annually

[a]*These estimates assume a shared source of reagents throughout the institution, four quarterly shipments, and monthly two-level QC of no more than two lots of cartridges/strips on a point-of-care reader or simple photometer where the chemistry of the test is in the disposable cartridge/strip, the cartridge/strip includes internal control for the chemical reaction, and the point-of-care reader includes built-in checks of its electronic performance. This example shows one approach that may be utilized to develop a QC Plan that saves analysis of external QC in lieu of internal control processes engineered by the manufacturer into the test system. The actual approach chosen will depend on the device, the manufacturer QC recommendations, the local regulations, and, most importantly, the medical director approval that the QC plan covers the primary risks in the testing process to appropriately meet medical needs.*

Using patient data in quality control procedures

Results from patient samples are used in four principal ways to support the QC processes in the laboratory:

- to verify consistency of patient results when changing lots of reagent or calibrator (discussed in a previous section),
- to perform a delta check with a previous patient result,
- to verify consistency of patient results when an analyte is measured using more than one instrument or method in a healthcare system, and
- to use a statistical process control scheme that tracks the mean (or median) of patient results to monitor method performance.

Delta check with a previous result for a patient

Some types of laboratory errors can be identified by comparing a patient's current test result to a previous result for the same analyte. This comparison is called a "delta check." Mislabeled samples and samples altered by dilution with IV fluid or collected with an incorrect preservative are examples of errors that can be detected using delta checks. The time period between two patient results must be small enough that a significant physiological change in the analyte is unlikely for a delta check to be effective to identify mislabeled or improperly collected samples. This limitation restricts the analytes that can be monitored with a delta check to those with small within-individual variability. The difference between results used for the delta parameter must be sufficiently large to avoid excessive numbers of false alerts; however, as the difference threshold becomes larger, the number of potential problems missed also increases. Kazmierczak [15] has reviewed delta check and other patient data-based QC procedures. CLSI guideline EP33 describes various approaches to using delta checks for quality management [16].

Verify consistency between more than one instrument or method

Another common use of patient results in a QC process is to verify consistency of patient results when an analyte is measured using more than one measuring system within the same health system. Good laboratory practice requires that multiple measuring systems (whether identical or based on different method principles) for the same analyte be calibrated to give clinically equivalent patient results whenever possible. It may be necessary to modify the calibration settings of one measurement system to match another system's results. This strategy allows a common reference interval to be used, provides continuity of results between patient encounters at different locations, and avoids clinical confusion regarding interpretation of laboratory results.

One method or instrument should be chosen to represent the primary comparative method to which others will be adjusted to achieve equivalent results. The primary method should be chosen based on quality and reliability of results with consideration of calibration traceability to national or international standards, performance stability,

selectivity for the analyte, and influence of interfering substances. CLSI guideline EP31 suggests procedures to evaluate the equivalence of patient results within a healthcare system [17]. The number of samples to use when verifying equivalence of patient sample results can be as few as one or more depending on the frequency of the assessment, the magnitude of a difference that is clinically meaningful, and the imprecision of the measuring systems. The laboratory will need to establish the frequency of evaluation and number of samples based on the stability of the methods, the frequency of reagent and calibrator lot changes, and the clinical requirements of the health system. Common practices include splitting one or more individual patient samples, or a small pool from several samples, on a weekly basis for high-volume methods, and a monthly, quarterly, or semiannual basis for lower volume or very stable methods. A larger number of samples are recommended if the monitoring is performed less frequently. When establishing interpretation criteria, the laboratory needs to consider the limited statistical power for the number of results available and utilize trend-monitoring techniques. For example, if one patient sample is tested weekly, the evaluation criteria for agreement will need to allow for the expected imprecision of the methods and rely more on trends to identify a measuring system that may be performing differently than others in the group.

Results for QC materials should not be used for verifying equivalence of patient sample results assayed using different measuring systems. Noncommutability of QC materials with patient specimens is a common occurrence. Even when more than one measuring system from the same manufacturer is used, QC results may show differences between different reagent lots and instrument models.

Using patient data for statistical process control

For a sufficiently large number of patient results, the mean (or median) value can be stable enough to be used as an indicator of measuring system consistency over time. This approach can be used on a periodic basis by extracting data for a time period, for example, 1 month, calculating the mean and SD for the distribution of results, and comparing one time period to another to determine if any changes have occurred. Automated approaches have been described to determine the mean (or median) for groups of sequential patient results for use as a process control parameter. These methods are called "average of normals" (AON) or "moving average" techniques and are suitable for use in higher volume assays in chemistry and hematology [18]. In general, these approaches evaluate sequential patient results over time intervals such as several hours to one or more days. The patients may be grouped by age, gender, and ethnicity to obtain homogeneous subgroups. Results may be trimmed to remove extreme values, so the remaining results are more reflective of patients without disease conditions that influence a particular analyte. However, trimming may remove results that are indicative of a calibration drift reducing the effectiveness of the procedure. The median is minimally influenced by extreme values and is recommended when the computer system can make the calculation. The mean (or preferably median) for a group of results is tracked to monitor trends in method performance using statistical procedures such as cumulative sum or exponentially weighted moving average [5,8]. These approaches can be a useful supplement to traditional surrogate QC materials when monitoring a single measuring system's stability and the calibration uniformity between multiple measuring systems in one or several distributed laboratories. Patient result-based algorithms are useful for high-volume settings but have not been widely adopted due to lack of consensus guidelines for their use and lack of computer support from instrument manufacturers and laboratory information systems.

Proficiency testing

Evaluation of method performance by an external entity is referred to as PT or external quality assessment (EQA). PT/EQA allows a laboratory to verify that its results are consistent with other laboratories using the same or similar measuring systems for an analyte and thus confirms the measuring system is being used correctly [19]. PT/EQA providers circulate a set of samples among a group of laboratories. Each laboratory includes the PT/EQA samples along with patient specimens in the usual assay process. The results for the PT/EQA samples are reported to the PT/EQA provider for evaluation.

Because PT/EQA samples are typically not commutable with native patient specimens, each reported result is compared to the mean result from all laboratories using the same measuring system or closely related ones (called a "peer group"). The report also includes the SD for the peer group distribution of results, the number of laboratories in the peer group, and the SDI or z-score, which expresses the reported result as the number of SD from the mean value. The limits of acceptability may be established by regulation or by the PT/EQA provider for analytes without regulatory criteria. The evaluation criteria are defined as a number of SD, a fixed percent, or a fixed concentration from the peer group mean value. Many QC manufacturers also provide a data analysis service that calculates summary statistics for measuring system peer groups similar to those for PT/EQA providers. This type of interlaboratory QC data allows a laboratory

to verify that it is producing results that are consistent with those of other labs using the same measuring system.

PT/EQA is designed to evaluate the total error of a single measurement. The acceptability limits for PT/EQA include laboratory bias and imprecision components plus other error components that are unique to PT/EQA samples such as between laboratory variation in calibration, variable matrix interaction with different lots of reagent within a peer group, filling imprecision of the PT/EQA material vials, stability variability in storage, shipping, and after reconstitution or opening of PT/EQC materials in the laboratory. Consequently, the acceptability limits for PT/EQA specimens are frequently larger than what might be expected for clinically acceptable total error with patient specimens.

PT/EQA is not offered for some analytes either because the test is new to the clinical laboratory or the analyte stability makes it difficult to include in a PT/EQA material. In these situations, the laboratory should use an alternate approach to verify acceptable performance of the method. CLSI guideline GP27 provides several approaches to assess measuring system performance when formal PT/EQA is not available [20].

Noncommutability of proficiency testing/external quality assessment materials and peer group grading

PT/EQA results are usually evaluated through peer group comparison because of the noncommutability of the materials typically used as PT/EQA specimens. Matrix interference is unique to each combination of measuring system and processed PT/EQA sample (just as for QC or reference materials) and can be quantitatively different for different method, instrument, and material combinations for the same analyte. Several investigations have reported >60% incidence of noncommutable materials for different analytes [1,11].

Fig. 5.6 illustrates the effect of noncommutable materials on interpretation of PT/EQA results. In this example, pooled native patient sera and PT/EQA samples were assayed by the DuPont Dimension analyzer and by the Abell−Kendall reference method for cholesterol [21]. The patient samples showed excellent agreement between the two methods (average bias = 0.2%). However, the PT/EQA materials had a large negative bias (−9.5%) with this measuring system that was due to noncommutability between the PT/EQA samples and the patient specimens with the Dimension measuring system. If the apparent accuracy of the routine method was compared against the reference method based on PT/EQA results, and the calibration was erroneously adjusted, the results for patient specimens would then be incorrect. PT/EQA results were still valuable to judge the performance of

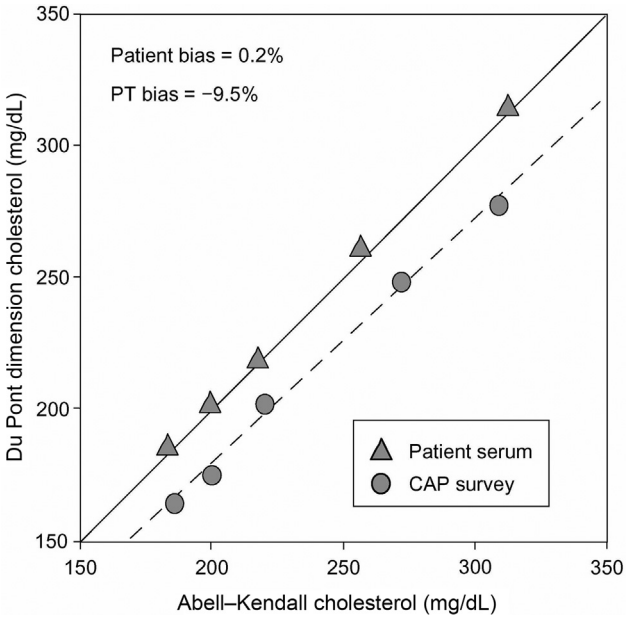


FIGURE 5.6 Example of noncommutability between proficiency testing materials and native patient sample pools for a specific method. *Adapted with permission from H.K. Naito, Y.S. Kwak, J.L. Hartfield, J.K. Park, E.M. Travers, G.L. Myers, et al. Matrix effects on proficiency testing materials: impact on accuracy of cholesterol measurement in laboratories in the nation's largest hospital system. Arch. Pathol. Lab. Med. 117 (1993) 345−351.*

laboratories using the Dimension method because the matrix interference was uniform within this peer group. Thus if an individual laboratory's results agreed with those of the peer group, the individual laboratory could conclude the measuring system was performing as expected. However, the accuracy of the Dimension method for patient specimens could not be evaluated based on the PT/EQA results.

Reporting proficiency testing/external quality assessment results when one method is adjusted to agree with another method

Good laboratory practice that is consistent with CLIA Regulations Section 493.1281 recommends adjusting the calibration of different measuring systems for the same analyte, within a health system, so the results for patient specimens are comparable irrespective of analytical method. In this situation, PT/EQA results must be reported correctly so they can be properly evaluated against the appropriate peer group. The peer group target value reflects the method calibration established by the measuring system manufacturer. For an individual laboratory's method with adjusted calibration, PT/EQA results must be reported to the PT/EQA provider with any user-applied calibration adjustment removed so the reported result can be compared to the manufacturer's calibration. The most convenient way to remove a calibration

adjustment is to first assay the PT/EQA samples with the calibration adjustment applied to the measuring system as would be the usual assay process for patient samples. After analysis, the PT/EQA results are corrected by mathematically removing the calibration adjustment. One should not recalibrate the instrument with a new set of calibrators for the purpose of analyzing PT/EQA specimens because this practice would violate regulations requiring the PT/EQA material to be assayed in the same manner as patient specimens.

For example, a laboratory has performed a native patient specimen comparison between Method A used in the main lab and Method B used in a satellite laboratory. Method B consistently generates results that are 10% higher than Method A, that is, a slope of 1.10 with negligible intercept. Method B can be adjusted to agree with Method A by automatically multiplying Method B's results by $1/1.10 = 0.91$ to lower the results by 10% before reporting (typically performed by the instrument's computer as a post measurement adjustment). When reporting PT/EQA results from Method B, it is necessary to remove the 0.91 factor to allow the reported result to be evaluated against the appropriate Method B peer group mean that will reflect the calibration established by the method manufacturer and allow the laboratory to assess its performance versus the peer group. Removing the 0.91 factor is accomplished by multiplying the reported PT/EQA result from Method B by the factor $1/0.91 = 1.10$ to increase its numeric value by 10%. The numeric result reported to the PT/EQA provider reflects the actual measured result using the manufacturer's recommended calibration settings. This process allows the PT/EQA sample to be assayed in the same manner as the patient specimens and the PT/EQA result to be appropriately evaluated by comparison to its peer group.

Interpretation of proficiency testing/external quality assessment results

When an unacceptable PT/EQA grade is received from a PT program, the measuring system should be investigated for possible causes and corrective action taken as necessary to correct any problems. Even when a PT/EQA grade is within acceptable limits, good laboratory practice recommends investigating any PT/EQA results that are greater than ~ 2.5 SDI from the peer group mean. When the SDI is 2.5, there is only a 0.6% probability of the result being within the expected distribution for the peer group; consequently, there is a reasonable probability of a measuring system problem that needs to be investigated. In addition, PT/EQA results that have been near the failure limit for more than one PT/EQA event, even if the results have passed the PT/EQA acceptance criteria, should initiate a review for systematic problems with the measuring system. These practices support identification of potential problems before they progress to more serious situations.

PT/EQA results are always received several weeks or more after the date of testing. Consequently, a review of QC, reagent, calibration, and maintenance records for the date of the testing and the preceding several weeks or months is necessary. If review of these records suggests a stable operating condition, and review of the PT/EQA material handling and documentation does not identify a cause for the erroneous PT/EQA result, it may be concluded the PT/EQA failure was a random event. The investigative steps, data reviewed, and conclusions from the review must be documented in a written report of the unacceptable PT/EQA result and reviewed by the laboratory director.

PT/EQA providers also include a summary report that includes the mean and SD/range for all the peer groups represented by the participants' results. Similar reports are available from interlaboratory QC programs. Summary reports are useful but must be interpreted with consideration of the limitations of noncommutable samples. The peer group mean and SD can be used to evaluate the uniformity of results between laboratories in the same peer group, to confirm an individual laboratory is using a measuring system in conformance to the manufacturer's specifications, and to evaluate the consistency of an individual laboratory's performance relative to the peer group over extended periods of time from one PT/EQA event to the next (trend monitoring). The summary information also allows evaluation of the imprecision of different types of measuring systems, and the number of users of each peer group reflects the market share for different measuring systems.

Because PT/EQA samples are commonly noncommutable with patient samples, PT/EQA summary reports cannot be used to compare: (1) the agreement or disagreement of patient results among different peer groups; (2) peer group means for different measuring systems to each other; or (3) an individual measuring system peer group mean to a value assigned by a reference method. Noncommutability also prevents any inference of the agreement for patient results among different measuring systems.

Accuracy-based proficiency testing/external quality assessment programs

PT/EQA providers offer programs that use commutable samples for some analytes. Commutable samples are typically prepared by pooling native clinical samples with minimal processing or additives to avoid any alteration of the sample matrix. To achieve samples with abnormal concentrations of analytes, donors can be identified with known

pathologic conditions, or blood and serum units from a general donor population can be prescreened for selected analytes. When commutable PT/EQA samples can be prepared, then the results from all laboratories reflect results that would be expected if native patient specimens were sent to each of the different laboratories. Thus agreement between laboratories (called "harmonization") and between different types of measuring systems can be correctly evaluated. The agreement between an individual laboratory result and a reference measurement result (accuracy), and the agreement between a measuring system mean and a reference method result (called trueness), can be evaluated when a reference method is available for an analyte.

For example, the College of American Pathologists Glycohemoglobin Survey has for many years used fresh, pooled whole blood from both normal and diabetic donors. The target values for the pooled blood are assigned by reference measurement procedures for hemoglobin A1c. In this survey, the accuracy of individual laboratory results, and the trueness of measuring system mean values, can be evaluated because the PT/EQA materials are commutable with native patient specimens. The measuring system trueness can be used by the respective manufacturers to monitor the effectiveness of their calibration processes.

PT/EQA programs have included commutable samples on occasion to evaluate individual laboratories and measuring systems groups for agreement with reference methods and for harmonization of results between laboratories and measuring systems groups [19]. Use of commutable materials adds substantial value to the information available from the results of PT/EQA programs.

References

[1] W.G. Miller, Specimen materials, target values and commutability for external quality assessment (proficiency testing) schemes, Clin. Chim. Acta. 327 (2003) 25–37.

[2] W.G. Miller, A. Erek, T.D. Cunningham, O. Oladipo, M.G. Scott, R.E. Johnson, Commutability limitations influence quality control results with different reagent lots, Clin. Chem. 57 (2011) 76–83.

[3] US Department of Health and Human Services, Medicare, Medicaid, and CLIA programs; laboratory requirements relating to quality systems and certain personnel qualifications; final rule. 42 CFR Part 493, Fed. Regist. 68 (2003) 3639–3714.

[4] CLSI, Statistical Quality Control for Quantitative Measurement Procedures: Principles and definitions: Approved Guideline C24-A4, Clinical and Laboratory Standards Institute, Wayne, PA, 2016.

[5] A.S. Neuhauer, The EWMA control chart: properties and comparison with other quality-control procedures by computer simulation, Clin. Chem. 43 (1997) 594–601.

[6] K. Linnet, Choosing quality-control systems to detect maximum clinically allowable errors, Clin. Chem. 35 (1989) 284–288.

[7] C.A. Parvin, A.M. Gronowski, Effect of analytical run length on quality-control (QC) performance and the QC planning process, Clin. Chem. 43 (1997) 2149–2154.

[8] T.P. Ryan, Statistical Methods for Quality Control, Wiley, New York, 1989, pp. 82–178.

[9] J.O. Westgard, P.L. Barry, M.R. Hunt, A multi-rule Shewhart chart for quality control in clinical chemistry, Clin. Chem. 27 (1981) 493–501.

[10] CLSI, User Evaluation of Between-Reagent Lot Variation: Approved Guideline EP26-A, Clinical and Laboratory Standards Institute, Wayne, PA, 2013.

[11] W.G. Miller, G.L. Myers, Commutability still matters, Clin. Chem. 59 (2013) 1291–1293.

[12] W.G. Miller, J.R. Tate, J.H. Barth, G.R. Jones, Harmonization: the sample, the measurement and the report, Ann. Lab. Med. 34 (2014) 187–197.

[13] CLSI, Laboratory Quality Control Based on Risk Management: Approved Guideline EP23-A, Clinical and Laboratory Standards Institute, Wayne, PA, 2011.

[14] Department of Health and Human Services. Centers for Medicare & Medicaid Services, IQCP announcement letter for CLIA CoC and PPM laboratories, September 30, 2013. <http://www.cms.gov/Regulations-and-Guidance/Legislation/CLIA/Individualized_Quality_Control_Plan_IQCP.html> (accessed April 2018).

[15] S.C. Kazmierczak, Laboratory quality control: using patient data to assess analytical performance, Clin. Chem. Lab. Med. 41 (2003) 617–627.

[16] CLSI, Use of Delta Checks in the Medical Laboratory: Approved Guideline EP33-A, Clinical and Laboratory Standards Institute, Wayne, PA, 2016.

[17] CLSI, Verification of Comparability of Patient Results Within One Healthcare System, Approved Guideline (Interim Revision) EP31-A-IR, Clinical and Laboratory Standards Institute, Wayne, PA, 2012.

[18] D. Ng, F.A. Polito, M.A. Cervinski, Optimization of a moving averages program using a simulated annealing algorithm: the goal is to monitor the process not the patients, Clin. Chem. 62 (2016) 1361–1371.

[19] W.G. Miller, G.R.D. Jones, G.L. Horowitz, C. Weykamp, Proficiency testing/external quality assessment: current challenges and future directions, Clin. Chem. 57 (2011) 1670–1680.

[20] CLSI, Using Proficiency Testing to Improve the Clinical Laboratory: Approved Guideline GP27-A3, Clinical and Laboratory Standards Institute, Wayne, PA, 2016.

[21] H.K. Naito, Y.S. Kwak, J.L. Hartfiel, J.K. Park, E.M. Travers, G. L. Myers, et al., Matrix effects on proficiency testing materials: impact on accuracy of cholesterol measurement in laboratories in the nation's largest hospital system, Arch. Pathol. Lab. Med. 117 (1993) 345–351.

Further reading

U.S. Department of Health and Human Services, Centers for Medicare and Medicaid Services. Appendix C, Survey procedures and interpretive guidelines for laboratories and laboratory services, Medicare, Medicaid, and CLIA programs; laboratory requirements relating to quality systems and certain personnel qualifications; final rule on January 24, 2003. CMS-2226-F: 42 CFR 493. <http://www.cms.hhs.gov/clia/appendc.asp> (accessed April 2018).

Self-assessment questions

1. What is the primary purpose of QC (statistical process control)?
   a. to determine that a method is adequate for clinical requirements
   b. to verify that a method is performing as expected for its stable operating condition
   c. to establish the uncertainty in a laboratory result
   d. to ensure that regulatory requirements are met
2. Which of the following criteria should be used to select materials to use for QC?
   a. analyte stability after opening
   b. analyte concentrations at low and high values relative to the analytical measurement range
   c. a matrix similar to that of clinical samples
   d. number of analytes in the same vial
3. How should the target value for a QC material be assigned?
   a. by assaying one bottle 10 times over 5 days
   b. by assaying 10 bottles on each of 10 days
   c. by assaying several bottles on each of 10 days while allowing the open-vial stability to age
   d. by using the manufacturer's assigned value
4. How is the expected variability for a stable measurement system established?
   a. by calculating the SD after a full month of QC results
   b. by calculating the cumulative SD over a 6-month interval
   c. by averaging the monthly SD over 6 months
   d. by calculating the cumulative SD over a long enough time period that all expected sources of variability have been included in the data
5. Which of the following sources of measurement variability need to be included in the SD used to evaluate a method's performance?
   a. instrument maintenance cycles
   b. reagent and calibrator lot changes
   c. QC material open-vial stability
   d. instrument parameters such as pipette stability, temperature stability, and dirt accumulation
6. Which of the following criteria should be used to select rules for evaluation of QC results?
   a. frequency of false alerts
   b. probability to detect significant bias
   c. probability to detect significant imprecision
   d. probability to detect trends in bias
7. Which of the following are considerations for how frequently to run QC samples?
   a. stability of the measurement system
   b. use of built-in controls
   c. clinical impact of reporting an incorrect result
   d. cost to repeat questionable results following a QC alert
8. Why might it be necessary to adjust QC target values following a reagent lot change?
   a. The calibration could be incorrect.
   b. The matrix interaction between the QC material and the new reagent could be different than with the old reagent.
   c. The method could have changed.
   d. The original target value could have been incorrect.
9. Why are QC or PT results not used to evaluate agreement between different methods?
   a. The concentrations are different in different samples.
   b. The within-peer group SD and CV include calibration variability between laboratories.
   c. The peer groups may not have enough participants for reliable statistics.
   d. The specimens used are frequently noncommutable with native clinical samples.
10. Which of the following components of the QC program are documented in the SOP and review logs?
    a. the specimen target values and SDs used for evaluation
    b. the results for each measurement
    c. troubleshooting performed to resolve a QC alert
    d. the acceptance criteria used to evaluate the QC results

Answers

1. b
2. a, b, c, d
3. c
4. d
5. a, b, c, d
6. a, b, c, d
7. a, b, c, d
8. b
9. d
10. a, b, c, d

Chapter 6

Laboratory calculations

Steven M. Truscott

Department of Pathology and Laboratory Medicine, Oakland University William Beaumont School of Medicine, Beaumont Health, Royal Oak, MI, United States

Learning objectives

After reviewing this chapter, the reader should be able to:

- Perform basic analytical calculations used in the clinical laboratory.
- Perform calculations commonly reported by the laboratory or done by clinicians to inform medical decisions.
- Perform basic method evaluation and laboratory management calculations.

Both clinicians and patients should be able to safely assume that laboratorians are correctly applying mathematical concepts in analytical testing, and that calculated values in patient reports are as reliable as measured values. Developing a lab-developed test (LDT) and assuring the quality of that test require the application of multiple types of calculations, from calibration and quality control to test system−specific operations and ongoing evaluation of method performance. Whether providing consulting service to clinicians or validating test systems, a laboratory director should know which calculations to use, how to assess the appropriate inputs, and how the calculations are done. He or she must understand the clinical impact of the calculations performed by the laboratory.

This chapter reviews important calculations performed in the clinical chemistry laboratory and focuses on the problem-solving aspects of these calculations. While mastery of the equations and mathematical operations is foundational, their application to the clinical laboratory is vital, which is why the calculation-focused portions of board examinations often take a case-based approach to probe one's ability to use the calculations in context. References are made to other chapters for more detailed descriptions of how each calculation is applied in various clinical or laboratory situations. Other laboratory mathematics texts or review texts offer additional opportunities for practice, from fundamental arithmetic to statistical methods [1−5].

Analytical calculations

Dimensional analysis

Dimensional analysis, or the factor-label method, is a useful approach for tracking conversion between units and relationships between physical quantities. Put simply, initial quantities must convert through *unit factors*, or ratios of equivalent or corresponding values, into desired quantities by "canceling units." This concept is demonstrated throughout the chapter for many types of relationships and provides a check on the accuracy of calculations describing chemical or physical relationships.

Example 6.1: *A serum creatinine result of 1.1 mg/dL is equivalent to ____ μmol/L? The molecular weight of creatinine is 113.1 g/mol.*

$$\begin{array}{l}\text{Initial quantity} \times \text{Unit factor} \times \text{Unit factor} \times \text{Unit factor} = \\ \text{Desired quantity}\end{array} \quad 1.1\ \text{m gdL} \times \frac{\text{mol}}{113.1\ \text{g}} \times \frac{1000\ \mu\text{mol}}{1\ \text{m mol}} \times \frac{10\ \text{dL}}{\text{L}} = \frac{97\ \mu\text{mol}}{\text{L}}$$

The same result can be obtained using tabulated factors for interconverting the conventional and SI units for analyte concentrations.

$$1.1\ \text{mg/dL} \times \frac{88.4\ \mu\text{mol}/L}{1\ \text{mg/dL}} = 97\ \mu\text{mol}/L$$

Example 6.2: *The concentration of free testosterone measured for a male patient is 218 pg/mL; total testosterone is 619 ng/dL. What fraction of the total testosterone is free?*

$$\frac{218\ \text{pg/mL}}{619\ \text{ng/dL}} \times \frac{\text{ng}}{1000\ \text{pg}} \times 100\ \text{mLdL} = 0.035 \text{ or } 3.5\%$$

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00006-5

Example 6.3: *A case report describes a patient with a serum urea concentration of 12.4 mmol/L. What is this patient's blood urea nitrogen (BUN) concentration in mg/dL? The molecular weight of urea is 60.1 g/mol.*

The BUN descriptor is historical; current methods measure serum samples and reported values are the concentration of nitrogen mass within urea. Converting between urea and urea nitrogen requires the mass ratio of nitrogen in the urea molecule.

$$12.4 \text{ m mol urea} \text{L} \times 60.1 \text{ gmol} \times \frac{\text{L}}{10 \text{ dL}} \times \frac{28.0 \text{ g N}}{60.1 \text{ g urea}}$$
$$= ? \text{ mg/dL BUN}$$

$$74.5 \text{ mg/dL urea} \times 0.466 \frac{\text{g N}}{\text{g urea}} = 34.7 \text{ mg/dL BUN}$$

Centrifugation

Centrifuges in the clinical laboratory are used to separate plasma or serum from the cellular components of the blood, or to prepare samples for analysis after extraction with solvents. When using different centrifuge equipment, the corresponding centrifugation speed must be determined for consistent separation. Relative centrifugal force (RCF, in units of "$\times$ *g*") is usually provided in published methods and by collection tube manufacturers. RCF is a function of the speed that the centrifuge rotates [*n*, in revolutions per minute (rpm)] and the radius [*r*, in centimeters (cm)] as measured from the center of the centrifuge rotor to the middle of the sample in the tube:

$$\text{RCF} = 1.118 \times 10^{-5} \times r \times n^2$$

Example 6.4: *Another lab in your hospital system has a centrifuge rotor with a 14-cm radius and prepares samples at 8000 rpm. To prepare samples with the same RCF using a 10-cm rotor, what speed in rpm should be used?*

$$\text{RCF}_1 = \text{RCF}_2$$

$$1.118 \times 10^{-5} \times r_1 \times n_1 2 = 1.118 \times 10^{-5} \times r_2 \times n_2 2$$
$$14 \text{ cm} \times (8000 \text{ rpm})^2 = 10 \text{ cm} \times n_2 2$$

$$n_2 = 9466 \text{ rpm}$$

Both centrifuges will operate at $1.118 \times 10^{-5} \times r \times n^2$, or $10{,}017 \times g$.

Concentration

The laboratorian should be adept at calculating reagent and analyte concentrations for common analytical methods and for the specific methods employed in his or her own laboratory. The most common units of concentration in the clinical laboratory are the percent weight/volume ($\%^{\text{w/v}}$) and molarity (mol/L or *M* in italics). Precisely defined, the percent weight/volume is grams of solute per 100.0 mL of total solution, and molarity is moles of solute per liter of total solution. Reagent concentrations may also be described in terms of mass per volume of total solution (e.g., mg/dL). Clinical results may be presented in grams or moles of analyte per volume of biological fluid (serum, plasma, whole blood, urine, or other fluid)—with conventional or SI prefixes as needed to be able to provide a manageable number to clinicians. The following examples review the preparation of simple solutions and interconversion of concentration units.

Example 6.5: *To prepare 0.100 L of a 100.0-mM aqueous stock solution of lactic acid, how many grams must be weighed and transferred to a volumetric flask?*

Start with the volume of solution that is needed and use the molecular weight of lactic acid as one of the unit factors.

$$0.100 \text{ L} \times 100.0 \text{ mmolL} \times \frac{\text{mol}}{1000 \text{ mmol}} \times 90.08 \text{ gmol} = 0.901 \text{ g}$$

Example 6.6: *An LDT method utilizes 1.5% NaCl. How many grams of NaCl are required to prepare 1.00 L? What is the molarity of this solution?*

Assume this is a percent weight/volume solution indicating the number of grams per 100.0 mL.

$$1.00 \text{ L} \times \frac{1.5 \text{ g}}{100.0 \text{ mL}} \times 1000 \text{ mLL} = 15.0 \text{ g}$$

Use the molecular weight of sodium chloride to convert into molarity.

$$\frac{15.0 \text{ g}}{\text{L}} \times \frac{\text{mol}}{58.44 \text{ g}} = 0.257 \text{ mol/L or } 257 \text{ mmol/L}$$

Example 6.7: *Zinc sulfate heptahydrate is used to make up a 5% solution of $ZnSO_4$ in deionized water. How many grams of the heptahydrate are needed to make 500.0 mL?*

One mole of a hydrated salt corresponds to one mole of an anhydrous salt; therefore the ratio of the two molecular weights becomes one of the unit factors.

$$500.0\ \text{mL} \times \frac{5.0\ \text{g}\ ZnSO_4}{100.0\ \text{mL}} \times \frac{287.56\ \text{g/mol}\ ZnSO_4 \bullet 7H_2O}{161.45\ \text{g/mol}\ ZnSO_4}$$
$$= 44.5\ \text{g}\ ZnSO_4 \bullet 7H_2O$$

Electrolyte concentrations may be expressed in terms of *equivalents* (Eq), where one equivalent is equal to one mole of charges [6]. To convert between moles and equivalents, it is helpful to remember the following pattern:

For monovalent ions, 1 mmol = 1 mEq
For divalent ions, 1 mmol = 2 mEq
For trivalent ions, 1 mmol = 3 mEq

Example 6.8: *A patient's serum magnesium concentration is reported as 0.8 mEq/L in another laboratory, but your lab's reference interval is 1.7–2.4 mg/dL. What is the patient's magnesium result in mmol/L? in mg/dL?*

Magnesium is a divalent cation, so 1 mmol Mg^{2+} corresponds to 2 mEq Mg^{2+}. The molarity can be determined first and then converted into mg/dL using the molecular weight of the magnesium ion.

$$\frac{0.8\ \text{mEq}\ Mg^{2+}}{\text{L}} \times \frac{1\ \text{mmol}}{2\ \text{mEq}} \times \frac{24.3\ \text{g}}{\text{mol}} \times \frac{\text{L}}{10\ \text{dL}}$$
$$= ?\ \text{mg/dL}\ Mg^{2+} \quad 0.4\ \text{m mol}\ Mg^{2+}\text{L} \times 24.3\ \text{gmol}$$
$$\times \frac{\text{L}}{10\ \text{dL}} = 1.0\ \text{mg/dL}\ Mg^{2+}$$

Other concentration-related calculations involving the concepts of normality, molality, and other percent solutions [e.g., percent weight/weight ($\%^{w/w}$) and percent volume/volume ($\%^{v/v}$)] are reviewed effectively elsewhere [4]. However, strong acids are usually labeled with percent weight/weight ($\%^{w/w}$), so the pertinent calculations should be reviewed briefly, since these reagents are frequently used in specialized methods and LDTs in clinical laboratories. Percent weight/weight ($\%^{w/w}$) refers to the number of grams of solute dissolved in 100.0 g of total solution. Specific gravity (SG) of the strong acid solution is also given on the chemical label. The SG, a unitless entity, is the ratio of the density of the solution to the density of water, which approaches 1.00 g/mL depending on the temperature. At room temperature, the SG of a solution is approximately the density of the solution in g/mL.

Example 6.9: *How many milliliters of concentrated HCl solution (37.0%$^{w/w}$, SG 1.19) are needed to prepare 500 mL of 0.200 M HCl?*

Start with the volume to be prepared. The desired molar concentration, molecular weight of HCl, percent weight/weight, and density (SG) are used as unit factors.

$$500\,\text{mL} \times \frac{0.200\,\text{mol HCl}}{\text{L}}$$
$$\times \left(\frac{36.46\,\text{g HCl}}{\text{mol HCl}} \times \frac{100.0\,\text{g sol'n}}{37.0\,\text{g HCl}} \times \frac{\text{mL}}{1.19\,\text{g sol'n}} \times \frac{\text{L}}{1000\,\text{mL}} \right)$$
$$= 8.28\,\text{mL}$$

Note that the unit factors indicated in parentheses can be used to determine the molarity of the concentrated HCl solution. One might find the molarity first, then the calculation simplifies to either of the following:

$$500\ \text{mL} \times 0.200\ \text{mol HClL} \times \left(\frac{\text{L}}{12.08\ \text{mol HCl}} \right) = 8.28\ \text{mL}$$

$$500\ \text{mL} \times \frac{0.200\ \text{M HCl}}{12.08\ \text{M HCl}} = 8.28\ \text{mL}$$

Dilutions

The last expression in Example 6.9 is consistent with an algebraic manipulation of the equation, $C_1V_1 = C_2V_2$. The product of concentration and volume of an initial solution (C_1V_1) equals the product of concentration and volume of a new solution (C_2V_2) after dilution or concentration, since the number of moles of solute does not change. In the context of this formula, the *dilution factor* is equal to V_2/V_1, the new total volume divided by the original volume. Laboratory procedures should be written with clear and consistent terminology for preparing dilutions. Potential confusion lies in differentiating among the *ratio to diluent*, the *dilution*, and the *dilution factor*.

Example 6.10: *A patient with bone metastases has a total serum alkaline phosphatase activity outside of the analytical measuring range. The lab procedure for manual dilutions calls for 50 μL of patient sample to be diluted with 150 μL of the approved diluent. Define the ratio to diluent, the dilution, and the dilution factor for this situation.*

Ratio to diluent: 1 to 3; 1 volume of sample (50 μL) to 3 volumes of diluent (3 μL × 50 μL)
Dilution: ¼, or 1 to 4; the ratio of sample volume (50 μL) to total volume of sample plus diluent (200 μL; note that a 1 to 4 dilution is not one part sample to four parts diluent).
Dilution factor: 4; the reciprocal of the dilution, or the total volume of sample plus diluent, divided by the original sample volume: 200 μL/50 μL = 4

When patient results are beyond a method's analytical measuring range and must be diluted, it is the *dilution factor* that should be used to calculate the final result after

reanalysis of the diluted sample. Many automated chemistry analyzers will perform on-board dilutions and calculate a final result. When further manual dilutions are required, the laboratorian should verify that both automated and manual dilution factors are applied correctly to obtain a final result. When serial dilutions are performed, the final dilution is the product of all individual dilutions. Dilution factors may be applied at the instrument, in middleware, or in the laboratory information system (LIS).

Example 6.11: *An automated immunoassay analyzer measures human chorionic gonadotropin (hCG) with an analytical measuring range of 2−1000 mIU/mL. Out-of-range samples undergo an automatic dilution of 1 to 100, and the dilution factor is applied automatically by the instrument. For a certain patient sample, the hCG result was still out of range after autodilution. A manual dilution was prepared using 50 μL of serum and 200 μL of the approved diluent. The hCG result provided by the instrument was 42,700 mIU/mL. What result should be reported to the clinician?*

The first dilution factor of 100 is already applied to the result by the instrument. The manual dilution was 1 to 5, so the final reported result should be

$$42,700\ \text{mIU/mL} \times 5 = 213,500\ \text{mIU/mL}.$$

Equilibrium calculations and pH

In the clinical laboratory, equilibrium calculations describe protein−ligand interactions, including antigen−antibody interactions, as well as acid−base equilibria. For any chemical reaction at equilibrium, the extent or state of equilibrium can be described mathematically using an equilibrium constant, K. For the reaction

$$aA + bB \rightleftarrows cC + dD$$

the equilibrium constant is determined from product and reactant concentrations and their stoichiometry. (Note that activity versus concentration of species in solution is not addressed here.)

$$K = \frac{[C]^c[D]^d}{[A]^a[B]^b}$$

Acid−base equilibria, both analytical and physiological, are among the most commonly encountered in the clinical laboratory. In a relatively dilute solution of weak acid, where changes in the molar concentration of water are insignificant, the equilibrium reaction can be represented as the dissociation of the acid (HA) into hydrogen ion and the conjugate base (A^-)

$$HA \rightleftarrows H^+ + A^-$$

and the equilibrium constant, in this case known as the acid dissociation constant, takes the form:

$$K_a = \frac{[H^+][A^-]}{[HA]}$$

Because hydrogen ion concentrations under physiological conditions are on the order of 1×10^{-7} mol/L, logarithms provide the most practical notation. The convention is to use pH for the negative logarithm of $[H^+]$ and to use pK_a for the negative logarithm of K_a.

$$pH = -\log [H^+]$$
$$pK_a = -\log K_a$$

It is worthwhile to review mathematical concepts involving logarithms [7] and how to use the "log" and "10^x" keys for your model of handheld calculator.

Buffer calculations

Buffer preparation: A buffer is an equilibrium solution, a mixture of a weak acid and its conjugate base (or a base and its conjugate acid), that resists pH changes when strong acid or base is added. For acid−base equilibrium and buffer calculations, one should remember the Henderson−Hasselbalch equation, which is derived from the expression for equilibrium dissociation constant of the acid:

$$pH = pK_a + \log \frac{[A^-]}{[HA]}$$

Buffers are completely specified by stating the volume needed, target pH, and total concentration [Y] of the specific acid/base pair, $[Y] = [A^-] + [HA]$. The ratio of $[A^-]$ / [HA] is determined using the Henderson−Hasselbalch equation based on the target pH.

Example 6.12: *Outline how to prepare 1.00 L of a pH 3.50, 0.200-M fluoride buffer: $pK_a = 3.20$, $K_a = 6.31 \times 10^{-4}$ for hydrofluoric acid.*

First, determine the $[A^-]$ / [HA] ratio using the Henderson−Hasselbalch equation:

$$pH = pK_a + \log \frac{[A^-]}{[HA]}$$

$$3.50 = 3.20 + \log \frac{[A^-]}{[HA]}$$

$$\frac{[A^-]}{[HA]} = 10^{(3.50-3.20)} = 2.00$$

Then, solve for $[A^-]$ and [HA], knowing that $[Y] = [A^-] + [HA] = 0.200\ M$.

$$[A^-] + [HA] = 0.200\ M$$
$$2[HA] + [HA] = 0.200\ M$$
$$[HA] = 0.067\ M$$
$$[A^-] = 0.133\ M$$

To prepare the buffer, add 0.133 mol of a fluoride salt and 0.067 mol of hydrofluoric acid to a 1.00 − L volumetric flask. Alternatively, add 0.200 mol of the fluoride salt and add 0.067 mol of a strong acid, or add 0.200 mol of hydrofluoric acid and add 0.133 mol of a strong base. Check and make minor adjustments to the pH using a pH meter, then dilute to 1.00 L with deionized water.

The carbonic acid buffer system: While buffer calculations involving carbon dioxide species are like other weak acid buffer calculations, there are significant complicating equilibria in the physiological setting, as described in detail in Chapter 35, Laboratory evaluation of kidney function. Carbonic acid (H_2CO_3) is a weak acid in equilibrium with bicarbonate (HCO_3^-), its conjugate base, and hydrogen ion (H^+). Yet carbonic acid is also in equilibrium with water and dissolved carbon dioxide gas, which can be exhaled to raise rapidly blood plasma pH. The key calculation for this buffer system is

$$\text{pH} = \text{pK}' + \log\frac{[HCO_3^-]}{\alpha \times \text{pCO}_2}$$

where pK′ is the apparent pK for carbonic acid, $[HCO_3^-]$ $[HCO_3^-]$ is the bicarbonate concentration in mmol/L, α is the solubility coefficient for CO_2 gas in mmol/L/mm Hg, and pCO_2 is the partial pressure in mm Hg of dissolved carbon dioxide gas. At the normal physiological temperature and ionic strength of human plasma, pK′ and α are 6.1 and 0.03, respectively; therefore:

$$\text{pH} = 6.1 + \log\frac{[HCO_3^-]}{0.03 \times \text{pCO}_2}$$

After measuring plasma pCO_2 and pH, the $[HCO_3^-]$ $[HCO_3^-]$ and total CO_2 (TCO_2) in mmol/L can be calculated. TCO_2 refers to the sum of bicarbonate, carbonic acid, and dissolved CO_2 gas:

$$TCO_2 = [HCO_3^-] + 0.03 \times \text{pCO}_2$$

This calculation is performed on blood gas analyzers and correlates well with total CO_2 methods on automated chemistry analyzers.

Example 6.13: *Blood gas results obtained for a patient with contraction alkalosis and hypokalemia include pH 7.52 and pCO_2 of 51-mm Hg. Determine bicarbonate and TCO_2 concentrations in mmol/L.*

$$\text{pH} = 6.1 + \log\frac{[HCO_3^-]}{0.03 \times \text{pCO}_2}$$
$$7.52 = 6.1 + \log\frac{[HCO_3^-]}{0.03 \times 51}$$
$$[HCO_3^-] = 0.03 \times 51 \times 10^{(7.52-6.1)}$$
$$[HCO_3^-] = 40\ \text{mmol/L}$$

To determine TCO_2, use this bicarbonate concentration.

$$TCO_2 = [HCO_3^-] + 0.03 \times \text{pCO}_2$$
$$TCO_2 = 40 + 0.03 \times 51 = 42\ \text{mmol/L}$$

Buffer capacity: In general, the ability of a buffer to resist pH changes, the buffer capacity, is the highest when the pK_a of the buffer system is near the intended pH and when near-equal concentrations of acid and conjugate base are present. (Interestingly, neither of these is the case for the carbonic acid buffer system, where the pK′ = 6.1 and the $[A^-]$/[HA] ratio is 20:1). Buffer capacity also increases with increasing total buffer concentration. However, depending on the analytical method, one may need to minimize the total buffer concentration. Using a variation of the Henderson−Hasselbalch equation, it is possible to determine how much total buffer concentration (conjugate base plus acid) is needed to resist a specified pH change induced by an added strong acid or base. Given initial pH (pH_i), as well as initial salt ($[A^-]_i$) and acid ($[HA]_i$) concentrations, a known concentration of added strong acid ([X]) will decrease the pH by δ. The added $[H^+]$ will react with the conjugate base, resulting in a higher concentration of undissociated weak acid, so $[A^-]$ will decrease by [X] and [HA] will increase by [X]:

$$\text{pH}_i - \delta = \text{pK}_a + \log\frac{[A^-]_i - [X]}{[HA]_i + [X]}$$

Adding a strong base raises the pH by δ and consumes the undissociated weak acid, resulting in a higher concentration of the conjugate base:

$$\text{pH}_i + \delta = \text{pK}_a + \log\frac{[A^-]_i + [X]}{[HA]_i - [X]}$$

Example 6.14: *An acetate buffer set to pH 4.6 must not allow pH to fall more than 0.2 pH units if up to 5.0 mM strong acid is added. What total buffer concentration is needed? (pK_a for acetic acid = 4.76.)*

Begin by calculating the $[A^-]_i$/$[HA]_i$ ratio from the pK_a and initial pH:

$$\begin{aligned}
\mathrm{pH_i} &= \mathrm{pKa} + \log\frac{[\mathrm{A^-}]_i}{[\mathrm{HA}]_i}\\
4.6 &= 4.76 + \log\frac{[\mathrm{A^-}]_i}{[\mathrm{HA}]_i}\\
\frac{[\mathrm{A^-}]_i}{[\mathrm{HA}]_i} &= 10^{(4.6-4.76)} = 0.692\\
[\mathrm{A^-}]_i &= 0.692\,[\mathrm{HA}]_i
\end{aligned}$$

Use the equation above for pH change upon addition of strong acid:

$$\begin{aligned}
&\mathrm{pH_i} - \delta = \mathrm{pKa} + \log\frac{[\mathrm{A^-}]_i - [\mathrm{X}]}{[\mathrm{HA}]_i + [\mathrm{X}]}\\
&4.6 - 0.2 = 4.76 + \log\frac{[\mathrm{A^-}]_i - 5.0\ \mathrm{m}M}{[\mathrm{HA}]_i + 5.0\ \mathrm{m}M}\\
&\frac{[\mathrm{A^-}]_i - 5.0\ \mathrm{m}M}{[\mathrm{HA}]_i + 5.0\ \mathrm{m}M} = 10^{(4.4-4.76)} = 0.437\\
&[\mathrm{A^-}]_i - 5.0\ \mathrm{m}M = 0.437([\mathrm{HA}]_i + 5.0\ \mathrm{m}M)\\
&0.692\,[\mathrm{HA}]_i - 5.0\ \mathrm{m}M = 0.437([\mathrm{HA}]_i + 5.0\ \mathrm{m}M)\\
&[\mathrm{HA}]_i = 28.2\ \mathrm{m}M\\
&[\mathrm{A^-}]_i = 19.5\ \mathrm{m}M
\end{aligned}$$

The total acetate buffer concentration should be at least 28.2 + 19.5 = 47.7 m*M*.

An alternative approach to buffer capacity calculations does not require the $\mathrm{pK_a}$ value if $[\mathrm{A^-}]_i$ and $[\mathrm{HA}]_i$ are known. Consider subtraction of the first two equations below:

$$\begin{aligned}
&\mathrm{pH_i} + \delta = \mathrm{pKa} + \log\frac{[\mathrm{A^-}]_i + [\mathrm{X}]}{[\mathrm{HA}]_i - [\mathrm{X}]}\\
&\mathrm{pH_i} = \mathrm{pKa} + \log\frac{[\mathrm{A^-}]_i}{[\mathrm{HA}]_i}\\
&\delta = \log\frac{[\mathrm{A^-}]_i + [\mathrm{X}]}{[\mathrm{HA}]_i - [\mathrm{X}]} - \log\frac{[\mathrm{A^-}]_i}{[\mathrm{HA}]_i}\\
&\delta = \log\frac{\dfrac{[\mathrm{A^-}]_i + [\mathrm{X}]}{[\mathrm{HA}]_i - [\mathrm{X}]}}{\dfrac{[\mathrm{A^-}]_i}{[\mathrm{HA}]_i}}\\
&10^{\delta} = \frac{\dfrac{[\mathrm{A^-}]_i + [\mathrm{X}]}{[\mathrm{HA}]_i - [\mathrm{X}]}}{\dfrac{[\mathrm{A^-}]_i}{[\mathrm{HA}]_i}}\\
&10^{\delta} \times \frac{[\mathrm{A^-}]_i}{[\mathrm{HA}]_i} = \frac{[\mathrm{A^-}]_i + [\mathrm{X}]}{[\mathrm{HA}]_i - [\mathrm{X}]}
\end{aligned}$$

This equation holds for the addition of a strong base, where the pH change is positive ($\delta > 0$). For the addition of strong acid, the pH change is negative ($\delta < 0$), and the equation becomes

$$10^{\delta} \times \frac{[\mathrm{A^-}]_i}{[\mathrm{HA}]_i} = \frac{[\mathrm{A^-}]_i - [\mathrm{X}]}{[\mathrm{HA}]_i + [\mathrm{X}]}$$

Example 6.15: *A phosphate buffer set to pH 7.4 is made with 190-mM Na_2HPO_4 and 50-mM NaH_2PO_4. What concentration of strong acid can be added before the pH drops by 0.5 pH units? (Three pK_a values apply to the triprotic phosphate buffer system, and they vary slightly in the literature; however, pK_a is not needed when the $[A^-]_i/[HA]_i$ ratio is known.)*

In this problem, $\delta = -0.5$, for a decrease in pH. Conjugate base will be converted into undissociated weak acid, so the correct calculation is

$$10^{\delta} \times \frac{[\mathrm{A^-}]_i}{[\mathrm{HA}]_i} = \frac{[\mathrm{A^-}]_i - [\mathrm{X}]}{[\mathrm{HA}]_i + [\mathrm{X}]}$$

$$10^{-0.5} \times \frac{190\ \mathrm{m}M}{50\ \mathrm{m}M} = \frac{190\ \mathrm{m}M - [\mathrm{X}]}{50\ \mathrm{m}M + [\mathrm{X}]}$$

$$(50\ \mathrm{m}M + [\mathrm{X}]) \times 1.202 = 190\ \mathrm{m}M - [\mathrm{X}]$$

$$[\mathrm{X}] = 59\ \mathrm{mM}\ \text{strong acid may be added}$$

Ionic strength

Ionic strength is a unitless quantity that accounts for the charge and concentration of all ions in a solution. It is used to estimate the difference between the activity (e.g., as measured by an ion-selective electrode) and the calculated concentration of the ion. The formula for calculating ionic strength (I) in any solution is

$$\mathrm{I} = \frac{1}{2}\sum c_i z_i^2$$

where c_i is the molar concentration of each individual ion, and z_i is the charge of each ion.

Example 6.16: *Determine the ionic strength of a 0.10 M solution of $MgCl_2$ at pH 7.0.*

The concentrations of each ion in this solution are

$$\begin{aligned}
[\mathrm{Mg^{2+}}] &= 0.10\ M\\
[\mathrm{Cl^-}] &= 0.20\ M\\
[\mathrm{H^+}] &= 1.0 \times 10^{-7}\ M\\
[\mathrm{OH^-}] &= 1.0 \times 10^{-7}\ M
\end{aligned}$$

Considering the hydrogen and hydroxide concentrations to be negligible in this case, use the concentrations and charges for $\mathrm{Mg^{2+}}$ and $\mathrm{Cl^-}$ to calculate *ionic strength*:

$$\mathrm{I} = \frac{1}{2}\left[(0.1)(+2)^2 + (0.2)(-1)^2\right] = 0.30$$

Photometry calculations

Photometric measurements are the basis of a majority of analytical systems in the clinical laboratory. The analytical requirements and biochemistry of absorbance spectrophotometry are described in Chapter 7, Spectrophotometry. Transmittance (T) and absorbance (A) of a solution are described mathematically in terms of either fraction or percentage of transmitted light intensity (I) relative to that of a blank solution (I_o).

$$\mathrm{T} = \mathrm{I}/\mathrm{I_o} \quad \text{or} \quad \mathrm{T} = \mathrm{I}/\mathrm{I_o} \times 100\%$$

$$\mathrm{A} = -\log \mathrm{T} \quad \text{or} \quad \mathrm{A} = 2 - \log \%\mathrm{T}$$

The Beer–Lambert law states that the absorbance of monochromatic light by a solution is directly proportional to the concentration of the absorbing species in the solution and to the path length and is expressed as

$$A^{\lambda} = \varepsilon^{\lambda} \times l \times c$$

where A^{λ} is the absorbance at wavelength λ, ε^{λ} is the *molar absorptivity* in liter/mol·cm at wavelength λ, l the length (cm) of the light path through the solution, and c the molar concentration of the solution. For a given analytical system, on measuring the absorbance of a solute over a range of concentrations, a plot of absorbance versus concentration is expected to form a straight line with an intercept of zero and a slope equal to the product of $\varepsilon^{\lambda} \times l$. The practical application of the Beer–Lambert law is to find the value for molar absorptivity ε from the slope of a plot of calibrator absorbances and from which the concentration of unknown samples can be determined from observed absorbance readings. Automated analyzers will record changes in absorbance over time in enzymatic reactions that are monitored continuously.

Using an extension of Beer's Law, the concentrations of mixtures of absorbing species can be determined. Thus it is possible to report multiple analyte concentrations from the same patient sample or to compensate mathematically for interfering substances that may be present at analytical wavelengths. For example, in a mixture with three absorbing compounds x_1, x_2, and x_3, the total absorbance at a single wavelength λ_1 would be expressed as

$$\mathrm{A}^{\lambda_1}_{\mathrm{total}} = \mathrm{A}^{\lambda_1}_{\mathrm{x_1}} + \mathrm{A}^{\lambda_1}_{\mathrm{x_2}} + \mathrm{A}^{\lambda_1}_{\mathrm{x_3}} = \varepsilon^{\lambda_1}_{\mathrm{x_1}} l c_{\mathrm{x_1}} + \varepsilon^{\lambda_1}_{\mathrm{x_2}} l c_{\mathrm{x_2}} + \varepsilon^{\lambda_1}_{\mathrm{x_3}} l c_{\mathrm{x_3}}$$

Absorbance measurements at multiple wavelengths (absorbance spectra) enable deconvolution of the contributions of each compound, and individual concentrations can be determined. Modern cooximeters can measure absorbance at >100 wavelengths simultaneously using diode arrays.

Electrophoresis

The migration of charged particles through liquid under an applied electric field is influenced by several factors (e.g., charge, size, and shape of solute; viscosity of the medium; ionic strength of the buffer; and strength of the electric field); see Chapter 8, Chromatography and electrophoresis. The electrophoretic rate of migration or migration velocity (v) is equal to the distance traveled (l, cm) per time of migration (t, s) and can be calculated as a function of the electrophoretic mobility (μ, $\mathrm{cm^2/Vs}$) and the field strength (E/d, V/cm). In electrophoretic theory, the electrophoretic mobility is found to be directly proportional to charge of the solute (Q) and inversely proportional to solute radius (r) and viscosity of the medium (η) [8].

$$v = \frac{l}{t} = \mu\frac{E}{d} = \left(\frac{Q}{6\pi r\eta}\right)\frac{E}{d}$$

For a given electrophoretic system, the electrophoretic mobility is the characteristic of the individual solute and can be used to estimate the migration rate when applied voltage is changed.

Example 6.17: *A voltage of 250 V is applied across a 5-cm gel, and a serum protein migrates 3 cm in 8 minutes. What is the electrophoretic mobility (μ)? How long will it take for this protein to migrate 3 cm if the voltage is reduced to 200 V?*

$$v = \frac{l}{t} = \mu\frac{E}{d}$$

$$\mu = \frac{l}{t} \times \frac{d}{E} = \frac{3\ \mathrm{cm}}{8\ \mathrm{min} \times 60\ \mathrm{s/min}} \times \frac{5\ \mathrm{cm}}{250\ \mathrm{V}}$$

$$= 12.5 \times 10^{-5}\ \mathrm{cm^2/Vs}$$

One "mobility unit" is defined as $10^{-5}\ \mathrm{cm^2/Vs}$, so the answer can also be stated as 12.5 mobility units. To find the new migration time, solve for t and substitute the new voltage:

$$t = \frac{l}{\mu} \times \frac{d}{E} = \frac{3\ \mathrm{cm}}{12.5 \times 10^{-5}\ \mathrm{cm^2/Vs}} \times \frac{5\ \mathrm{cm}}{200\ \mathrm{V}}$$

$$= 600\ \mathrm{s\ or}\ 10\ \mathrm{min}$$

The migration rate is proportional to the applied voltage.

Electrochemistry

The most common electrochemical method in the clinical laboratory is potentiometry, which can be applied to the measurement of pH, as well as sodium, potassium, chloride, lithium, and magnesium ion concentrations. As described in Chapter 9, Electrochemistry, the Nernst equation can be derived and simplified to a form that is most readily used for determining the electrical cell potential as a function of ion activity (presuming the cell has been designed to be selective for the ion of interest). This equation is:

$$E_{cell} = E^{\circ} + \left(0.0592/n\right) \times \log a$$

where E_{cell} represents the measured voltage for the cell, E° is the standard potential, n is the charge of the ion, and a is the activity of the ion of interest. A further simplifying assumption can be made, where the molar concentration is used for the ion activity. Note that the measurement of cell potential must be quite accurate and precise, as the change in potential is only 0.0592 V for a 10-fold change in ion concentration, for ions with a plus or minus one charge. For the divalent cations (calcium, magnesium) where the value for n is 2, the sensitivity is lowered to 0.0592/2 or 0.0298 V for a 10-fold concentration change.

Coulometric titration is the most precise method available for quantification of chloride ions and is used for sweat chloride measurements (see Chapter 27, Testing in alternative matrices: CSF and body fluid analysis). A constant current of silver ions reacts with the dissolved chloride ions in a sample to form insoluble silver chloride. The reaction is timed to detect a threshold concentration of excess silver ions. Using a blank sample, the time for threshold detection in the absence of chloride ions can be measured and subtracted. Total charge (Q) transferred is proportional to the number of moles of chloride present in the sample. The relationship is described by Faraday's law

$$Q = It = znF$$

where Q is charge in coulombs, I is current in amperes, t is time in seconds, z is number of charges transferred (per reaction), n is the number of moles reacted, and F is the Faraday constant (96,485 coulombs per mole). Remember that an ampere is one coulomb per second, A = C/s or C = A s.

Example 6.18: *A 100-μL sample has a titration time of 17.5 seconds on a chloridometer operating at 48 mA. The blank time was 1.2 seconds. What is the chloride concentration in this sample in mmol/L?*

Subtract the blank time to determine how long chloride ions from the patient sample were reacting with the silver ion current: 17.5−1.2 = 16.3 s. Solve for number of moles and divide by the volume of the sample to determine chloride concentration.

$$Q = It = znF$$

$$n = It/zF$$

$$n = \frac{48 \text{ m C/s} \times 16.3 \text{ s}}{1 \times 96485 \text{ C /mol}} = 8.1 \times 10^{-3} \text{ mmol Cl}^{-}$$

$$\frac{8.1 \times 10^{-3} \text{ mmol Cl}^{-}}{100 \text{ μL}} \times \frac{10^{6} \text{ μL}}{\text{L}} = 81 \text{ mmol/}L \text{ Cl}^{-}$$

This concentration of chloride in a sweat sample is consistent with a diagnosis of cystic fibrosis.

Enzyme kinetics

A single-substrate enzymatic reaction can be described by the equation

$$E + S \rightleftarrows \text{ES} \rightarrow E + P$$

where E is the enzyme, S is the substrate, ES is the enzyme−substrate complex, and P is the product of the reaction. The Michaelis−Menten equation describes the relationship between the rate of enzymatic activity and the initial substrate concentration:

$$v = \frac{V_{max}[S]}{K_m + [S]}$$

where v is the enzyme-catalyzed reaction rate, V_{max} the maximum reaction rate, $[S]$ the substrate concentration, and K_m the Michaelis−Menten constant, which is equal to $[S]$ at ½ V_{max}. When enzymes are used as analytical reagents to determine *substrate concentration* in a patient sample, the enzyme is added in large excess of the expected substrate concentration. This establishes the first-order kinetics in which the $[S]$ is very low compared with the K_m, and the Michaelis−Menten equation reduces to

$$v = \left(\frac{V_{max}}{K_m}\right)[S] = k\,[S].$$

The first-order rate constant is k, and the reaction rate is directly proportional to the substrate concentration. When the intent is to measure *enzyme activity* in a patient sample, the substrate is added in large excess of the expected enzyme concentration. This establishes zero-order kinetics, in which $[S]$ is much larger than the K_m, and the Michaelis−Menten equation reduces to $v = V_{max}$. The enzyme activity is measured directly as the rate of substrate depletion or product formation over time.

One of the first steps in developing an enzyme assay is to determine the kinetic constants K_m and V_{max} for the assay conditions. For the most part, this is now done by importing kinetic data into software applications. However, a Lineweaver–Burk plot provides a straightforward example of the basis of these calculations. The Lineweaver–Burk plot is a linear transformation of the hyperbolic Michaelis–Menten curve. The reaction velocity (e.g., μmol/min) is measured at multiple substrate concentrations (e.g., μmol/L), and reciprocal velocity ($1/v$) is plotted versus reciprocal substrate concentration ($1/[S]$). The Lineweaver–Burk equation takes the basic form of a linear equation ($y = mx + b$). The slope equals K_m/V_{max}, and x- and y-intercepts correspond to $-1/K_m$ and $1/V_{max}$, respectively.

$$\frac{1}{v} = \left(\frac{K_m}{V_{max}}\right)\left(\frac{1}{[S]}\right) + \frac{1}{V_{max}}$$

Example 6.19: *Determine K_m and V_{max} from the following kinetic data for enzyme activity.*

v (μmol/min)	$[S]$ (μmol/L)
8	5
15	10
28	20
48	40
73	80
94	160
100	240
104	320
106	400

Using a spreadsheet, calculate $1/v$ and $1/[S]$ for each data pair. Regular linear regression of $1/v$ versus $1/[S]$ gives a slope (m) and intercept (b):

$$m = 0.5894 = \frac{K_m}{V_{max}};\ K_m = 83\ \mu mol/L$$

$$b = 0.0071 = \frac{1}{V_{max}};\ V_{max} = 141\ \mu mol/min$$

Under conditions of zero-order kinetics, one can determine enzyme concentration using photometric data and the molar extinction coefficient ε for the optically active component of the reaction, as determined by calibration. The following equation accounts for the dilution of a patient sample (V_{sample}) into the total reaction volume (V_{total}) and utilizes both the Beer–Lambert law and the definition of an international unit of enzyme activity (IU = μmol $_{substrate}$/min). The change in absorbance ΔA is determined for a reaction time t in a reaction vessel of path length l.

$$[E] = \left(\frac{\Delta A/\varepsilon l}{t}\right)\left(\frac{10^6\ \mu mol}{mol}\right)\left(\frac{V_{total}}{V_{sample}}\right)$$

Example 6.20: *Aspartate aminotransferase (AST) activity is determined photometrically using a coupled reaction (which is not rate-limiting) that oxidizes NADH at a rate proportional to AST activity. In this test system, the cuvette path length is 0.2 cm and ε for NADH is 6.22×10^3 L mol^{-1} cm^{-1}. When 10 μL of patient serum is added to 50 μL of reagents at 37°C, the absorbance change (ΔA) is −0.032 over a monitored reaction time of 5 minutes. What is the AST concentration in IU/L?*

The absorbance change is negative, because NADH absorbance decreases when it is oxidized to NAD^+ at 340 nm. Using the equation above

$$[E] = \left(\frac{0.032\ /\ \left[\left(6.22 \times 10^3\ \text{L mol}^{-1}\ \text{cm}^{-1}\right)(\ 0.2\ \text{cm}\ )\right]}{5\ \text{min}}\right)$$

$$\left(10^6\ \mu\text{molmol}\right)\left(\frac{60\ \mu\text{L}}{10\ \mu\text{L}}\right)[E] = 31\ \mu\text{mol/min/L} = 31\ \text{IU/L}$$

Chromatography

Calculations related to chromatography are applicable to many mass spectrometry-based LDTs (see Chapter 8, Chromatography and electrophoresis, and Chapter 10, Mass spectrometry). This section reviews basic calculations describing chromatographic peaks and the extent of their separation by column chromatography. The retention factor (k') quantifies how long a compound is retained on a column, the retention time (t_r), relative to the column void time (t_o). For isocratic separations, k' is independent of the flow rate of the mobile phase. This factor can also be calculated in terms of volumes.

$$k' = \frac{t_r - t_o}{t_o} = \frac{v_r - v_o}{v_o}$$

The selectivity factor (α) describes the relative retention of two compounds. For this formula, the retention time for the compound that elutes later is given the subscript of 2.

$$\alpha = \frac{k'_2}{k'_1} = \frac{t_{r2} - t_0}{t_{r1} - t_0} = \frac{v_{r2} - v_o}{v_{r1} - v_o}$$

Calculating the chromatographic resolution (R_s) of two peaks accounts for both retention and the widths of two chromatographic peaks. It is the difference in retention times (or volumes), divided by the average of the baseline widths of the two peaks.

$$R_s = \frac{t_{r2} - t_{r1}}{\left(\frac{(w_{b1} + w_{b2})}{2}\right)}$$

Clinical calculations

Physicians from the various medical specialties will perform calculations based on numeric results provided by the laboratory in order to interpret the data and make medical decisions. In some cases, the laboratory performs these calculations in the LIS and reports the calculated results automatically. The College of American Pathologists requires verification of calculated values that appear with patient results, biennially or whenever a change in the test systems or information systems could affect the calculations. Examples of several clinical calculations are given below. Note that these kinds of calculations have assumptions and limitations that should be considered and must be interpreted in the context of the whole clinical picture.

Electrolytes

Anion gap: This calculation is used in the differential diagnosis of metabolic acidosis (see Chapter 36, Blood gas and critical care testing). The anion gap (AG) is the difference in mmol/L between the measured cations and the measured anions in serum and is traditionally calculated as

$$AG = [Na^+] - [Cl^-] - [HCO_3^-]$$

although some laboratories include potassium.

$$AG = [Na^+] + [K^+] - [Cl^-] - [HCO_3^-]$$

The anion gap is 8−16 mmol/L in healthy individuals (12−20 mmol/L, if K^+ is included). An elevated anion gap indicates the presence of unmeasured anions (e.g., lactate, acetoacetate, and β-hydroxybutyrate) in the serum that may be involved in the acidosis. As shown in Example 6.29, the imprecision of the anion gap calculation is considerable, because the variances of all individual measurands contribute. In fact, for any of the following clinical calculations, the imprecision will increase with the number of and with the individual imprecisions associated with component measurands used for the calculation.

Delta ratio: For an uncomplicated metabolic acidosis, one would expect the increase in unmeasured anions to be reflected in a decrease in the base, HCO_3^-. The increase is calculated as the anion gap minus 12 (an average normal anion gap). The decrease is calculated as 24 (an average normal bicarbonate concentration) minus the patient's measured $[HCO_3^-]$. The ratio of these differences is the delta ratio.

$$\frac{AG - 12}{24 - [HCO_3^-]}$$

Significant deviation from a delta ratio of 1 suggests the presence of a mixed acid−base disorder (again see Chapter 36, Blood gas and critical care testing).

Example 6.21: *A diabetic patient in ketoacidosis had the following results for a basic metabolic panel. Calculate the anion gap and delta ratio. What type of additional acid−base disorder is suggested by these results?*

Na^+	139	mmol/L
K^+	4.7	mmol/L
Cl^-	92	mmol/L
HCO_3^-	16	mmol/L
Glucose	378	mg/dL
BUN	29	mg/dL
Creatinine	1.7	mg/dL

$$\text{Anion gap} = [Na^+] - [Cl^-] - [HCO_3^-] = 139\text{-}92\text{-}16 = 31 \text{ mmol/L}$$

$$\text{Delta ratio} = \frac{AG - 12}{24 - [HCO_3^-]} = \frac{31 - 12}{24 - 16} = \frac{19}{8} = 2.4$$

The increase of 19 mmol/L in the anion gap is more than twice the decrease in bicarbonate concentration. There is an additional metabolic alkalosis, which in this case was due to loss of acid by vomiting associated with a viral gastroenteritis.

Osmolality and osmolal gap

Osmolality is the number of moles of dissolved particles (charged or uncharged solutes) per kilogram of solution. Most laboratories use freezing point depression to measure this colligative property in serum or urine. The primary clinical utility is to detect osmotically active substances that may not be rapidly measured in the laboratory, such as toxic alcohols or mannitol. Osmolality measurements are also used to assess the kidney's ability to concentrate urine. It is possible to predict serum osmolality based on measurements of serum sodium, BUN, and glucose:

$$\text{Calculated serum osmolality} = 2 \times [Na^+](\text{mmol/L}) + \frac{[\text{Glucose}](\text{mg/dL})}{18} + \frac{[\text{BUN}](\text{mg/dL})}{2.8}$$

If ethanol is measured in the serum, its osmolality can be added to the calculated value as [EtOH] (in mg/dL)/4.6. If the difference between the measured and calculated osmolality (osmolal gap) is more than 10 mOsm/kg for a particular serum sample, other osmotically active

substance(s) must be present (see Chapter 37, Water and electrolyte balance).

Renal function calculations

Renal function testing has been a challenge for the clinical laboratory, as the kidney has substantial reserve capacity, such that substances cleared from the blood do not show a marked increase in concentration until the reserve is nearly exhausted. Various methods have been combined with patient demographics in calculations that estimate renal function in terms of glomerular filtration rate (GFR).

Renal clearance: The classic method for estimating GFR is creatinine clearance, which is calculated from the serum creatinine concentration (C_s, mg/dL) and the urine creatinine concentration (C_u, mg/dL) of a urine sample collected over a known time interval (t, minute), usually 24 hours, and the measured urine volume (V_u, mL).

$$\text{Creatinine clearance}\quad (\text{mL/min}) = \frac{C_u}{C_s} \times \frac{V_u}{t}$$

The test relies on creatinine as an endogenous marker for GFR and requires a reliable urine collection, which is difficult to obtain, especially outside of the hospital environment.

GFR estimating equations: Laboratorians should be familiar with equations used to calculate estimated GFR (eGFR). Each is based on serum creatinine measurement and does not require urine collection. Cockcroft–Gault, Modification of Diet in Renal Disease study (MDRD), and Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equations, in order of increasing accuracy relative to measured GFR, pertain to estimates in adults. The 2009 modification of the Schwartz equation, also known as the "Bedside Schwartz" equation, is appropriate for children ages 1–18 years old [9]. Note that even the currently recommended CKD-EPI equation still provides estimates that differ significantly from measured GFR in many adult patients. All of these equations are limited by the inherent physiology of creatinine. Details pertaining to the clinical application of these equations are described more fully in Chapter 35, Laboratory evaluation of kidney function. The Cockcroft–Gault equation, first described in 1976, is

$$\text{eGFR}_{CG} = \frac{(140\text{-age}) \times \text{weight}}{72 \times [S_{Cr}]} \times 0.85\ [\text{if female}]$$

where age is in years, weight in kilograms, and $[S_{Cr}]$ is the serum creatinine concentration in mg/dL. The 0.85 constant is applied if the patient is female; similar constants are applied in the MDRD and CKD-EPI equations where applicable to the patient's sex and race. Cockcroft–Gault estimates are not adjusted for body surface area (BSA) and are inconsistent with current automated creatinine methods, which have been standardized to the isotope dilution mass spectrometry (IDMS) reference method.

The MDRD equation derives from the 1999 Modification of Diet in Renal Disease study:

$$\text{eGFR}_{MDRD} = 175 \times ([S_{Cr}])^{-1.154} \times \text{age}^{-0.203} \times 0.742\ [\text{if female}] \times 1.210\ [\text{if African American}]$$

Again, $[S_{Cr}]$ is in mg/dL and age in years. The initial constant of 175 is appropriate for current, IDMS-standardized creatinine assays.

The CKD-EPI equation was published in 2009 by the Chronic Kidney Disease Epidemiology Collaboration:

$$\text{eGFR}_{CKD\text{-}EPI} = 141 \times \min([S_{Cr}]/\kappa, 1)^{\alpha} \times \max([S_{Cr}]/\kappa, 1)^{-1.209} \times 0.993^{\text{age}} \times 1.018\ [\text{if female}]\ \times 1.159\ [\text{if African American}]$$

- $[S_{Cr}]$ is IDMS-standardized serum creatinine in mg/dL.
- κ is 0.7 for females and 0.9 for males.
- α is −0.329 for females and −0.411 for males.
- Min indicates the minimum of $[S_{Cr}]/\kappa$ or 1.
- Max indicates the maximum of $[S_{Cr}]/\kappa$ or 1.

Further revisions of the CKD-EPI equation have been published, which incorporate serum cystatin C concentration [10], but these equations have not yet been widely adopted. Cystatin C is discussed further in Chapter 35, Laboratory evaluation of kidney function.

The modified Schwartz equation is at present the best creatinine-based GFR estimate for children and is also adjusted for the IDMS-traceable creatinine methods currently in use:

$$\text{eGFR}_{\text{modified Schwartz}} = \frac{0.41 \times \text{height}}{[S_{Cr}]}$$

The result of this calculation has units of mL/min/1.73 m^2, as for CKD-EPI; the height of the patient should be entered in centimeters (cm), and in this version of the equation, serum creatinine in mg/dL.

Except for the historical Cockcroft–Gault calculation, eGFR results are reported as normalized values, adjusted to the average adult BSA, which is why eGFR is reported in units of mL/minute/1.73 m^2. For patients with a significantly larger or smaller than average BSA, the adjustment may be removed to obtain an eGFR specific to the patient [11]. To do this, multiply the eGFR by the patient's BSA in m^2, then divide by 1.73:

$$\text{eGFR} = (\text{eGFR, mL/min/1.73m}^2) \times \text{BSA}_{\text{patient}}, \text{m}^2/1.73$$

Online calculators and original references for the GFR estimating equations are available from the National Kidney Foundation [12] and the National Institute of Diabetes and Digestive and Kidney Diseases [13].

Example 6.22: *Determine the eGFR (CKD-EPI) for an African American female, age 57, whose serum creatinine is 1.30 mg/dL. BSA for this patient is 1.95 m^2.*

Enter the serum creatinine. For a female, κ is 0.7 and α is -0.329. For the second term, the minimum is 1, and for the third term, the maximum is 1.30/0.7 or 1.86. Enter the age in years as the exponent for 0.993. On most models of handheld calculator, the [y^x] key should be used to apply the exponent to the base, which key is not the same as the [EXP] key to program scientific notation ($\times 10^x$). Both of the constants for sex and race are applied to obtain the calculated eGFR.

$$\begin{aligned}
&\mathrm{eGFR_{CKD\text{-}EPI}} \\
&= 141 \times \min\left([S_{Cr}]/\kappa,\ 1\right)^{\alpha} \times \max\left([S_{Cr}]/\kappa,\ 1\right)^{-1.209} \times 0.993^{\mathrm{age}} \\
&\quad \times 1.018\ [\text{if female}] \times 1.159\ [\text{if African American}] \\
&= 141 \times \min\left(1.30/0.7, 1\right)^{-0.329} \times \max\left(1.30/0.7, 1\right)^{-1.209} \\
&\quad \times 0.993^{57} \times 1.018 \times 1.159 \\
&= 141 \times (1)^{-0.329} \times (1.86)^{-1.209} \times 0.993^{57} \times 1.018 \times 1.159 \\
&= 52.6\ \mathrm{mL/min/1.73\ m^2}
\end{aligned}$$

To remove the normalization for BSA, use the equation above.

$$\mathrm{eGFR} = 52.6\ \mathrm{mL/min/1.73\ m^2} \times 1.95\ \mathrm{m^2}/1.73 = 59.3\ \mathrm{mL/min}$$

Fractional excretion of sodium: This calculation requires sodium and creatinine measurements in both blood and urine. It is used in patients with acute kidney disease to determine the fraction of sodium excreted by the kidneys, using creatinine as a marker to adjust for urine concentration.

$$\mathrm{Fe_{Na}} = \frac{[\mathrm{Na^+}]_{\mathrm{urine}} \,/\, [\mathrm{Na^+}]_{\mathrm{serum}}}{[\mathrm{Cr}]_{\mathrm{urine}} \,/\, [\mathrm{Cr}]_{\mathrm{serum}}} \times 100\%$$

$\mathrm{Fe_{Na}} < 1\%$ is an indication of prerenal disease, or decreased blood flow to the kidneys. $\mathrm{Fe_{Na}} > 1\%$ indicates renal damage. As with each clinical calculation, there are certain limitations in the application of $\mathrm{Fe_{Na}}$, and the calculation should be interpreted in clinical context (see Chapter 37, Water and electrolyte balance).

Iron calculations

Serum iron and iron-binding capacity calculations are helpful in the differential diagnosis of iron deficiency or overload conditions (see Chapter 47, Trace elements). Serum iron concentration should be interpreted in the context of the total iron-binding capacity (TIBC) of transferrin and percent saturation of transferrin (Tf); the clinical laboratory should report these results together. TIBC is measured directly by some methods, or TIBC can be calculated by first measuring unbound iron-binding capacity (UIBC) of transferrin, then adding the serum iron concentration:

$$\mathrm{TIBC}\ (\mu\mathrm{g/dL}) = \mathrm{UIBC}\ (\mu\mathrm{g/dL}) + [\mathrm{Fe_{serum}}]\ (\mu\mathrm{g/dL})$$

The percent saturation of transferrin calculation is

$$\text{Tf saturation}\ (\%) = \frac{[\mathrm{Fe_{serum}}](\mu\mathrm{g/dL})}{\mathrm{TIBC}\ (\mu\mathrm{g/dL})} \times 100\%$$

Example 6.23: *Determine the percent transferrin saturation for a sample with serum iron concentration of 90 μg/dL and a UIBC of 330 μg/dL.*

$$\mathrm{TIBC} = 330\ \mu\mathrm{g/dL} + 90\ \mu\mathrm{g/dL} = 420\ \mu\mathrm{g/dL}$$

$$\text{Tf saturation}\ (\%) = \frac{90\ \mu\mathrm{g/dL}}{420\ \mu\mathrm{g/dL}} \times 100\% = 21.4\%$$

To determine the percent saturation of transferrin on a molar basis after direct measurement of serum transferrin, an alternate calculation can be performed based on the molar ratio of iron to transferrin. Note that two moles of iron bind to one mole of transferrin and that the constant 0.673 is applicable when units of measure for iron and transferrin are conventional:

$$\begin{aligned}
&\text{Molar Tf saturation}\ (\%) \\
&= \frac{[\mathrm{Fe_{serum}}]\ (\mu\mathrm{g/dL})}{[\mathrm{Tf_{serum}}]\ (\mathrm{mg/dL})} \times \frac{\mathrm{mg}}{1000\ \mu\mathrm{g}} \\
&\quad \times \frac{\text{mol Fe}}{55.84\ \mathrm{g}} \times 75{,}200\ \text{gmol Tf} \\
&\quad \times \frac{1\ \text{mol Tf}}{2\ \text{mol Fe}} \times 100\% = \frac{[\mathrm{Fe_{serum}}]}{[\mathrm{Tf_{serum}}]} \times 0.673 \times 100\%
\end{aligned}$$

Transferrin is the major iron transport protein in plasma; therefore the percent saturation calculations should be relatively close, since other serum proteins bind iron to a lesser extent than transferrin.

Example 6.24: *Determine the molar percent transferrin saturation for a serum iron concentration of 90 μg/dL and transferrin of 310 mg/dL.*

$$\text{Molar Tf saturation}\ (\%) = \frac{90\ \mu\mathrm{g/dL}}{310\ \mathrm{mg/dL}} \times 0.673 \times 100\% = 19.5\%$$

Lipid calculations

Chapter 28, Lipoprotein disorders, describes the pathophysiology and biochemical measurements of cholesterol and serum lipoproteins. Cholesterol present in low-density

lipoprotein (LDLC) is fundamental to assessment of cardiovascular risk. LDLC can be measured directly or calculated using the Friedewald equation, which requires three other lipid measurements, each in mg/dL, taken on the same fasting sample:

$$\text{LDLC} = \text{TC-HDLC-}\frac{\text{TG}}{5}$$

Note that the overall error in the LDLC calculation depends on the error in measurements of total cholesterol (TC), high-density lipoprotein cholesterol (HDLC), as well as triglycerides (TG). For this relationship to be valid, chylomicrons must be absent, as is the case in normal fasting samples, and TG must be less than 400 mg/dL. The serum triglyceride result is divided by a factor of 5 to estimate the cholesterol found in very-low-density lipoprotein particles (VLDLC). Recently, adjustable factors for estimating VLDLC have been proposed, but consensus guidelines have not yet recommended these calculations for routine use [14].

The non-HDLC fraction, which includes cholesterol in all lipid fractions thought to contribute to cardiovascular risk, is not dependent on fasting or triglyceride levels and should be calculated automatically in the LIS and reported with lipid panels. Non-HDLC is simply the difference between TC and HDLC in mg/dL:

$$\text{Non-HDLC} = \text{TC-HDLC}$$

For patients with highly elevated levels of lipoprotein (a) [Lp(a)], LDLC may appear to be refractory to therapy, since Lp(a) cholesterol is a subfraction of the LDLC. In such cases, recent guidelines have suggested calculation in mg/dL of Lp(a)-corrected LDLC [14].

$$\text{Lp(a)-corrected LDLC} = \text{TC-HDLC-}\frac{\text{TG}}{5}\text{-[Lp(a)} \times 0.30\text{]}$$

This approach may help clinicians with treatment decisions, since Lp(a) is known not to respond to statin therapy.

Statistics

Statistical calculations are used extensively in the clinical laboratory to evaluate method performance. Spreadsheet programs and software packages are helpful with calculations, but the laboratorian must establish the quality of the input data, determine which statistical calculations are appropriate, and interpret the calculated statistics properly (see Chapter 2, Statistical methods in laboratory medicine). This section will review basic statistical operations used in the clinical laboratory that can readily be performed with a handheld calculator.

Mean

The most commonly used measure for the central tendency of a set of values, the mean ($\bar{x}$) is calculated by finding the sum of the data set and dividing by the number of values in the set. For large data sets, especially when values are repeated within the set, calculators that have a "statistics" mode can be used to enter the set more efficiently.

$$x = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example 6.25: *Determine the mean for the following data set: 63, 68, 69, 63, 65, 66, 65, 67, 61, 65.*

$$\bar{x} = \frac{652}{10} = 65.2$$

Median

The median is a measure of central tendency that is less susceptible to outliers. More often used for non-Gaussian data, it is independent of assumptions about the distribution pattern of the data set and requires only inspection of the set, once the set is arranged in numerical order. The arrangement can be either ascending or descending, and the middle value for the set is the median.

Example 6.26: *Determine the median for the following values: 63, 68, 69, 63, 65, 66, 65, 67, 61, 65.*

Arrange the set and find the middle value: 61, 63, 63, 65, **65, 65**, 66, 67, 68, 69. There is an even number of values in the set; in which case, the average of the two central values is the median. In this example, both central values are the same, 65, so no further calculation is needed.

Standard deviation

The "SD" or *s* statistic is used to describe the logical concept of the spread of values for a normally distributed (Gaussian) data set. While the range for a set also gives an idea of the spread, the SD is an improvement, in that all of the data enter into the calculation. While the concept is straightforward, the calculation of standard deviation (SD) is not. When the data set is a sample of the population, the formula is

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

where the difference between each of the individual values (x_i) and the mean ($\bar{x}$) is squared, the squared differences are summed, and this sum is divided by the number of values in the set, less one. The SD is the square root of the quotient and has the same units as the individual data points. The mean is determined before the SD calculation is done. When calculating the SD by hand, it is helpful to construct a table to track the individual values for the squares of the differences. In practice, the calculation of SD is performed in a software package, middleware, or the LIS and is most often used to evaluate method imprecision.

Example 6.27: *Determine the SD for the following data set: 63, 68, 69, 63, 65, 66, 65, 67, 61, 65.*

Creating a table as described above, or using a handheld calculator or spreadsheet program, the SD calculates to 2.4. Some calculators have an option of σ_n or σ_{n-1} for this calculation. The latter (sample SD) is appropriate for a data set that is a sample of a complete set, as is the case for precision studies.

Coefficient of variation

While the SD is usually used to describe precision, comparing SDs for two data sets can be complicated when the means or the units for each set differ. For example, a comparison of the SD for creatinine and BUN data sets would not determine which method is more precise, without knowing the means. To facilitate such comparisons, the coefficient of variation (CV) is calculated and reported as a percentage, using the equation

$$\mathrm{CV} = s\bar{x} \times 100\%$$

Example 6.28: *A control was tested 10 times to validate within-run precision for an FDA-approved method. Mean ($\bar{x}$) was 65.2 mg/dL, and SD (s) was 2.4 mg/dL. Determine the CV.*

$$\mathrm{CV} = \frac{2.4}{65.2} \times 100\% = 3.7\%$$

The CV is not presented in units of the assay, so the amount of variation can be readily compared with other methods or concentrations. However, the utility of this statistic is reduced as the mean approaches zero.

Variance

When combining independent components of measurement imprecision, it is not appropriate to add SDs. The variance, s^2, or square of the SD, is the additive statistic. Given three data sets x, y, and z, the total variance is calculated as the sum of squared SDs.

$$\left(s_{TOTAL}\right)^2 = {s_x}^2 + {s_y}^2 + {s_z}^2$$

This concept applies to estimates of uncertainty when a reported result is calculated from the sum of or the difference between independent measurements. Consider anion gap in the following example [15].

Example 6.29: *The SDs for sodium, potassium, chloride, and bicarbonate measurements are found to be 1.2, 0.1, 1.3, and 1.2 mmol/L, respectively. What is the imprecision in the anion gap?*

$$\begin{aligned} s^2_{\mathrm{AG}} &= s^2_{\mathrm{Na+}} + s^2_{\mathrm{K^+}} + s^2_{\mathrm{Cl^-}} + s^2_{\mathrm{HCO_3^-}} \\ s^2_{\mathrm{AG}} &= (1.2)^2 + (0.1)^2 + (1.3)^2 + (1.2)^2 \\ s^2_{\mathrm{AG}} &= 4.58 \\ s_{\mathrm{AG}} &= 2.14\ \mathrm{mmol/L} \end{aligned}$$

This is relatively imprecise, as calculated anion gap results would be expected to vary as much as $\pm 2s_{\mathrm{AG}}$ or ± 4 mmol/L (at the 95% confidence interval) on repeated measurements of the same patient sample.

If a calculated clinical result is a product or quotient of independent measurements, then the overall imprecision must be calculated from relative SDs, or CVs:

$$(\mathrm{CV_{TOTAL}})^2 = \mathrm{CV}_x^2 + \mathrm{CV}_y^2 + \mathrm{CV}_z^2$$

Standard error of the mean

Replicate measurements of control materials are actually samples taken from the larger population of possible measurements. For a parametric data set, how close is a mean of sample measurements ($\bar{x}$) to the true mean (μ) of all possible measurements? For this one computes the standard error of the mean (SEM), or s-subscript $\bar{x}$, calculated as

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where s is the SD of the sample set and n is the number of data points in the set. The more data collected, the smaller the standard error, and the more precise the estimate of μ. Standard error can be calculated for other statistics, not just the mean, and is used to calculate the confidence intervals based on tabulated t-values (see Chapter 2, Statistical methods in laboratory medicine). As an approximation, for a large data set, the true mean μ lies within ± 2 SEM from the sample mean $\bar{x}$ at the 95% confidence level.

Method evaluation

The statistical calculations described above, as well as several other calculations presented here, are used to evaluate the analytical and clinical performance of laboratory methods. The companion concepts for these calculations are discussed in more detail in Chapter 4, Method evaluation.

Accuracy: regression analysis of method comparison data

To evaluate bias in a quantitative method, the usual data treatment is to plot the results for the method being evaluated on the *Y*-axis versus the results from the same patient samples tested by a reference method (or best comparative method) on the *X*-axis of a scatter plot. Linear regression analysis determines the best fit line by providing the slope and the intercept, which indicate proportional bias and constant bias, respectively. An ideally accurate comparison has a slope of 1.00 and an intercept of zero. The regression slope and intercept are heavily influenced by outlier data at concentration extremes, so the data should be evenly distributed across measurable concentrations. How well the individual data points fit the line is given by the correlation coefficient, *r*, where the perfect fit of all the points gives an *r* value of 1.00. However, this value is not useful for determining bias; two methods may have excellent correlation but still have significant bias.

The laboratorian should be aware of the strengths and weaknesses of three types of linear regression. Regular linear regression minimizes differences in one direction, so it is easily calculated. However, this mathematical approach assumes no error (imprecision) in the *X*-direction (comparative method), and is therefore not often appropriate for real method evaluations. Deming regression accounts for imprecision in both methods, but is still affected when data sets are not evenly distributed. Passing−Bablok regression, a nonparametric calculation, makes no assumptions of data distribution. Method evaluation software will perform any of these regression types. However, for a small data set, the Passing−Bablok slope and intercept can be approximated with a basic calculator. The principle is to determine the slope and intercept for all combinations of *X*−*Y* data points in the data set and then determine the medians.

Example 6.30: *Determine the Passing−Bablok slope and intercept for the following comparison data set:*

	X	Y
1	*24*	*26*
2	*50*	*48*
3	*72*	*74*
4	*97*	*93*

Complete the data table below with the slope (change in *Y* divided by change in *X*) for each data pair and then calculate the intercept for each pair from the equation $b = Y - m\ X$, substituting the *X* and *Y* values from the larger of the two data sets in the pair.

Pair combination	ΔY	ΔX	Slope	Intercept
1−2	22	26	0.846	5.69
1−3	48	48	1.000	2.00
1−4	67	73	0.918	3.97
2−3	26	22	1.182	− 11.09
2−4	45	47	0.957	0.13
3−4	19	25	0.760	19.28
Median	**−**	**−**	**0.938**	**2.99**

The Passing−Bablok slope is 0.938 and intercept is 2.99. For comparison, Deming regression analysis gives a slope of 0.943, an intercept of 2.9, and a correlation coefficient of 0.997.

Accuracy: recovery studies

An evaluation method may be checked for accuracy by adding a known amount of analyte to a sample, then testing the sample to determine if the added amount is recovered. When a nonaqueous solvent must be used as a spiking solution, a blank with the same solvent should be incorporated into the study design to account for matrix differences. Since a recovery study requires manipulation of native samples, this assessment of accuracy may be considered secondary to a correlation study that includes a range of patient samples. The prepared samples may not match native samples for matrix differences, such as the presence of metabolites and protein binding. This limitation actually turns out to be useful when recovery studies can demonstrate that a method provides comparable results within allowable error for different sample matrices, for example, in a nonstandard body fluid. The fundamental calculation for any recovery study is to determine accurately the expected value for the analyte concentration in the prepared samples, allowing for dilution and for analyte initially present. This is an application of the equation used above in the "Dilutions" section:

$$C_1V_1 = C_2V_2$$

The product of the initial concentration and volume (C_1V_1) for the analyte should equal the product of the final concentration and volume (C_2V_2) after the sample has been manipulated. If the analyte is present in both the original sample and a spiking sample, then the C_1V_1 for both samples must be added and this sum is divided by the new total volume (V_2) to determine the expected concentration (C_2).

Example 6.31: *A laboratory is validating a glucose method in the testing of pleural fluid using serum/fluid admixtures. The glucose concentration in a serum pool is 103 mg/dL, and the glucose concentration in a pleural fluid sample is 164 mg/dL. An admixture of 1.2 mL of serum and 0.8 mL of fluid has a measured glucose concentration of 124 mg/dL. Calculate the percent recovery of glucose in the admixture.*

First calculate the expected concentration of glucose in the admixture.

$$C_{1(\text{serum})}V_{1(\text{serum})} + C_{1(\text{fluid})}V_{1(\text{fluid})} = C_{2(\text{admixture})}V_{2(\text{admixture})}$$

$$103\ \text{mg/dL} \times 1.2\ \text{mL} + 164\ \text{mg/dL} \times 0.8\ \text{mL} = C_{2(\text{admixture})} \times 2.0\ \text{mL}$$

$$C_{2(\text{admixture})} = 127\ \text{mg/dL}$$

Then calculate the recovery as a percentage of the expected concentration.

$$\frac{124\ \text{mg/dL}}{127\ \text{mg/dL}} \times 100\% = 97.6\%$$

Depending on matrix effects on a given method and the assay precision, it is possible to recover greater than 100%.

Precision: repeatability and within-lab precision studies

For the initial assessment of method precision, a within-run study of replicate measurements, usually $n = 20$, will provide the data to determine the mean, SD, and CV for a single run (see the "Statistics" section above). If this is acceptable, then a more extensive within-lab precision study is carried out over several days to address additional sources of variability that will be encountered in routine use of the method. Analysis of variance calculations performed by statistical software requires a data set of several smaller ($n = 2$ or 3) within-run tests of relevant control materials, two runs per day, each day for 20 days. The various contributions to the overall imprecision are calculated from this data set, and total imprecision is estimated for routine testing of quality control materials.

Total error

Error in an analytical method consists of both bias (inaccuracy) and imprecision, and a clinician has no way of determining to what extent either contributes to the error in a given patient result. Therefore the laboratorian must combine both sources of error and assure that the *total error* (TE) does not exceed allowable error. At the 95% confidence interval, TE is calculated in units of the method as

$$\text{TE} = \text{Bias} + 1.96 \times \text{SD}$$

TE is calculated at specified analyte concentrations, usually at medical decision levels (MDL). SD is determined from precision studies at the MDLs, and bias is calculated from the regression statistics (slope and intercept) determined from method comparisons. The analyte concentration corresponding to the MDL is entered as the X value in the regression equation ($Y = m\,X + b$), and the bias is the difference between Y and X. That is,

$$\text{Bias}_{\text{MDL}} = (mX_{\text{MDL}} + b)\text{-}X_{\text{MDL}}$$

Example 6.32: *Determine the TE for a method at the MDL of 200 mg/dL. Regression statistics for the method comparison study gave a best fit of $Y = 1.034X - 2.2$ and the within-lab precision studies indicated an SD of 1.2 mg/dL at 200 mg/dL.*

$$\text{TE} = \text{Bias} + 1.96 \times \text{SD}$$

$$\text{TE} = [(mX_{\text{MDL}} + b) - X_{\text{MDL}}] + 1.96 \times \text{SD}$$

$$\text{TE} = [(1.034\ (200) - 2.2) - 200] + 1.96 \times 1.2$$

$$\text{TE} = 4.6 + 1.96 \times 1.2 = 7.0\ \text{mg/dL}$$

TE is compared with allowable error for the method. Allowable error is determined by the laboratory director, who considers primarily clinical but also regulatory requirements for method performance characteristics.

Detection limits

Calculating detection limits for an evaluation method is done by measuring imprecision at very low concentrations over multiple days. So again, the mathematical approach aligns with the "Statistics" section. The lowest quantitative estimate of detection limit is *limit of blank* (LoB). A zero concentration sample is measured in replicate and the SD is calculated from these data. The LoB may be estimated as 2 × SD of blank replicates; thus 2.5% of blank sample measurements would still fall above this threshold. The *limit of detection* (LoD) is quantitatively greater than LoB. Replicate samples of known very low concentration are measured and, as an approximation, 2 × SD of these replicates is added to the LoB. For LDTs, the most important estimate of the detection limit is the *limit of quantitation* (LoQ), which defines functional sensitivity and the lower limit of the measuring interval, below which quantitative values should not be reported.

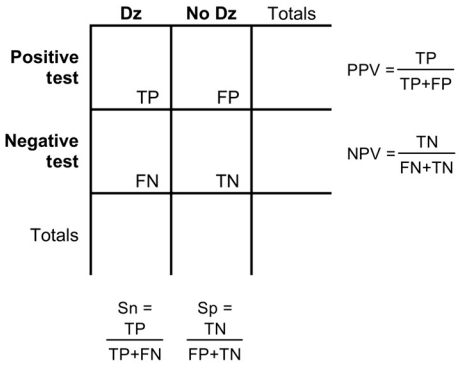


FIGURE 6.1 Example template for predictive value theory calculations. To calculate sensitivity, specificity, or predictive values, divide the true tests in the row or column by the total for that row or column. Prevalence of the disease in a given cohort or population equals the total of the first column divided by the total number of individuals. *Dz*, Individuals with disease; *FN*, false negative; *FP*, false positive; *No Dz*, individuals without disease; *NPV*, negative predictive value; *PPV*, positive predictive value; *Sn*, sensitivity; *Sp*, specificity; *TN*, true negative; *TP*, true positive.

The LoQ is the lowest value that can be measured within stated acceptability limits for accuracy and precision. Repeated measurements at low concentrations define the analyte concentration at which the method is accurate (e.g., $<10\%$ bias from a comparative method) and precise (e.g., $<20\%$ CV). Detailed descriptions of the LoB, LoD, and LoQ calculations, both parametric and nonparametric, are described in CLSI EP17 [16].

Clinical performance of laboratory tests: predictive value theory

Either the manufacturer (for FDA-approved systems) or the laboratory (for LDTs) must validate that a diagnostic test is clinically useful. The calculations of predictive value theory evaluate how well a method classifies patients into diagnostic categories. Test results produced by the evaluation method are compared either to a gold standard method performed on the same patient cohort or to a reliable clinical diagnosis assigned to each member of the cohort. Depending on quantitative cutoffs selected, each test result is designated as true positive (TP), true negative (TN), false positive (FP), or false negative (FN). Here, we review how to calculate sensitivity (Sn), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), and diagnostic efficiency. These quantities are usually calculated as percentages. To avoid confusion, it is helpful to keep a consistent approach to these types of problems by drawing a simple grid the same way for each scenario (Fig. 6.1). Then, depending on what information is available, expect the starting values for the grid to differ. The reader is encouraged to draw and fill in a grid by hand for each example below.

Example 6.33: *In a prostate cancer study, 5000 men were screened by measuring serum PSA concentration. Two hundred and fifty of the men had a PSA above 4.0 ng/mL and were given a biopsy. Two hundred out of the 250 had a negative biopsy result, and 50 had positive biopsy findings. Follow-up studies on the 4750 men who had PSA ≤ 4.0 ng/mL showed that 75 had prostate cancer at the time of screening. Calculate diagnostic performance statistics for this PSA test in this study group, including Sn, Sp, PPV, and NPV, and diagnostic efficiency.*

A total of 5000 tests were performed, 250 of which were positive and 4750 of which were negative at this PSA cutoff. These are entered as the totals for each row. The biopsy result is the gold standard, so we know that 50 of the men with positive tests had the disease (TP) and 200 did not (FP). Seventy-five of the PSA-negatives actually did have the disease (FN). Thus far, solely from the given data, the grid should look like this:

	Dz	No Dz	Totals	
Positive test	50	200	250	PPV = %
Negative test	75		4750	NPV = %
Totals			5000	
	Sn = %	Sp = %		

Now complete the grid by calculating the missing values. FNs plus TNs must add up to the total of all negative tests ($75 + TN = 4750$); therefore $TN = 4675$. A total of 125 subjects had prostate cancer (Dz) and 4875 were cancer-free (No Dz). Enter these totals for the columns. The grid is now complete, and the diagnostic statistics are calculated from the true results divided by the total in each row or column. Diagnostic efficiency is the sum of true results divided by the total number of tests: $(50 + 4675) / 5000 = 0.945$ or 94.5%. Note that Sn cannot be calculated from testing healthy subjects, nor can Sp from testing diseased/unhealthy subjects.

	Dz	No Dz	Totals	
Positive test	50	200	250	PPV = 20%
Negative test	75	4675	4750	NPV = 98%
Totals	125	4875	5000	
	Sn = 40%	Sp = 96%		

Sn and Sp can be considered unchanging characteristics of the test, independent of the population being tested, as long as the quantitative cutoff is not changed. (Lowering the cutoff value will increase Sn, but at the expense of decreased Sp.) Predictive values, however, are highly dependent on the prevalence of the disease in the population being studied. In the cohort in Example 6.33, the prevalence of prostate cancer was 125/5000, the number with disease divided by total subjects, or 2.5%. If a broader population of men is considered, in which prostate cancer prevalence is lower, and the same test is used for screening, then the predictive values can be recalculated.

Example 6.34: *Given the Sn and Sp of the PSA screening test in the previous example, what would the PPV and NPV be for a population in which the prevalence of prostate cancer was 0.25%?*

Draw a grid as before, and consider what is known about this new population. To obtain the predictive values, one can assume a convenient number (e.g., 100,000) for the total population. Based on the prevalence in this population, 0.25% of 100,000 or 250 will have the disease:

	Dz	No Dz	Totals	
Positive test				PPV = ?
Negative test				NPV = ?
Totals	250		100,000	
	Sn = 40%	Sp = 96%		

Individuals without disease can be found by subtraction: 250 + No Dz = 100,000; therefore the total of No Dz = 99,750. All of the other entries in the grid can be calculated from the Sn and Sp of the test. Forty percent of diseased individuals will test positive: 0.40 × 250 = 100 TP; the rest (150) will be FN. Ninety-six percent of healthy individuals will test negative: 0.96 × 99,750 = 95,760 TN; the rest (3990) will be FP. Row totals are calculated, and PPV and NPV are calculated from the number of true results divided by totals in each row:

	Dz	No Dz	Totals	
Positive test	100	3990	4090	PPV = 2.4%
Negative test	150	95,760	95,910	NPV = 99.8%
Totals	250	99,750	100,000	
	Sn = 40%	Sp = 96%		

In this broader population, the PPV is much lower. Even though the Sn of the test is still 40%, the probability that a patient with a positive test actually has the disease is only 2.4%.

Another mathematical approach to describe Sn and Sp is to calculate likelihood ratios (LR). These are defined in terms of the likelihood of obtaining a positive or negative test result *in a person with the disease*. The LR answers the question, "Assuming disease is present, what is the likelihood of a positive result (LR +) or negative result (LR −)?" This is distinct from PPV and NPV, which predict presence or absence of disease, given a test result. The positive LR is calculated as LR + = Sn/(1 − Sp), and the negative LR is calculated as LR − = (1 − Sn)/Sp. A higher LR + or lower LR − indicates a more useful test, and as with Sn and Sp, LRs are not affected by prevalence of the disease.

Laboratory management

Management of a clinical laboratory involves not only analytical and medical considerations, but also multiple aspects of business and healthcare administration [17]. Laboratorians should be familiar with several key financial calculations.

Budget justification

Creating and monitoring a budget do not require complicated mathematics per se, but the volume of calculations may be daunting depending on the complexity of budget categories. Budgets describe the expected revenue and expenditures for a given timeframe, according to the institutional fiscal year. Budgets for an upcoming fiscal year are justified based on historical trends, expected contractual changes, marketing studies, supply costs, and staffing needs. After detailed review, the upcoming year's budget figures may be established based on an absolute or

percent change from the current year, with clear written justification for each budget category. Annual budgets may be broken down to monthly revenue and expenditures to account for seasonal or short-term variations. Significant increases or decreases from expected budget figures, or variances, must be accounted for. Corrective action may be needed, for example, when the laboratory is able to control supply purchases or finalize delinquent payments.

Full-time equivalent

The full-time equivalent (FTE) represents the typical workload of 40 hours per week and 52 weeks per year. FTE expenses include both salary and benefits, often approximated as an additional 20%−30% of the salary. In laboratory management, the work unit of FTEs is used to allocate the workforce to various sections of the lab, individual instrument systems, or workstations. In such allocations, full work efficiency is not assumed, as factors such as vacations, shifts, illness, breaks, training, etc. need to be considered. Finally, actual staffing requires whole people, so calculated FTEs may need to be rounded up to the next integer, unless part-time employment can be used.

Cost per reportable result

The apparent cost of a lab test can vary by over 10-fold, depending on how it is calculated. The lowest cost per test comes from the simple estimate of the reagent cost to perform one test, but that value by itself has little utility in laboratory cost management. The *cost per reportable* result, or CPR, comes from a more complex calculation, but provides a basis for making important decisions such as whether an assay should be performed in-house or by a reference laboratory. The CPR is calculated by summing all costs allocated to the test (total cost) and dividing by the number of test results reported out to a patient record (reportable results). The number of reportable results in the denominator excludes QC or calibration samples, test repeats, or losses due to outdated reagents. However, the costs associated with these excluded tests are included in the numerator, along with other expenses. Note that a reportable result may not be equivalent to a *billable test*, which may consist of a panel or test group billed as a single item, even though it may contain several reportable results. The numerator of the total cost estimate includes both *direct* (reagents, QC materials, instrument costs, operator costs, etc.) and *indirect* costs (hospital overhead, clerical staff, and management), as allocated to the assay. An alternative approach is to consider costs as either *fixed* or *variable*. Here, costs that change with the number of tests reported (e.g., reagents and operator cost) are variable, while fixed costs are independent of the number of reported tests (instrument costs, calibration, and overhead). By dividing costs into fixed and variable, changes in CPR can be readily made as a function of changes in the test volume. Note that the CPR will go down as the test volume goes up, since the fixed costs are being allocated over a larger number of reportable results. This calculation can be used to determine the "break-even" test volume, the threshold number of tests where in-house costs would equal the send out cost for a given period of time. Beyond this point, all tests performed in-house would represent savings for the lab.

Example 6.35: *Determine the cost per reportable result for a test system with the following costs: instrument annual maintenance contract of $35,000, and annual cost of $65,000 based on 5-year depreciation. Operator costs are 0.2 FTE at $80,000/year per FTE. Reagent costs are $1.24 per test. Total annual tests are estimated at 14,000, of which 23% are not reportable (QC, calibrators, external proficiency samples, etc.).*

The total cost per year is the sum of instrument maintenance ($35,000), instrument cost ($65,000), allocated operator cost (0.2 × $80,000), and reagent cost ($1.24 × 14,000). This sum is divided by the number of reportable results per year (77% of 14,000).

$$\begin{aligned}\text{CPR} &= \frac{\text{Total cost}}{\text{Reportable results}} \\ &= \frac{\$35,000 + \$65,000 + (0.2 \times \$80,000) + (\$1.24 \times 14,000)}{0.77 \times 14,000} \\ &= \frac{\$133,360}{10,780} = \$12.37\end{aligned}$$

Example 6.36: *Plasma vitamin E costs $32 for each sample sent to a reference laboratory. To perform the test in-house, the total fixed costs, including a new HPLC detector, would be $22,000. Variable costs, including reagents, supplies, and a portion of an FTE, are estimated at $7.75 per reportable test. What would be considered a break-even test volume?*

Determining the break-even test volume (v) requires breaking down the total cost per reportable to separate components of fixed and variable costs as they approach the cost of sending out the test to the reference laboratory.

$$\text{Fixed} + \text{Variable} = \text{Sendout cost}$$

$$\$22,000 + v \times (\$7.75 \text{ per test}) = v \times (\$32 \text{ per test})$$

$$v = \frac{\$22,000}{(\$32\text{-}\$7.75)} = 907 \text{ tests}$$

Capital equipment justification

The laboratory can improve the quality of service to patients, decrease the turnaround times, and avoid the higher sendout costs by investing in improved instrumentation for in-house testing. Large capital expenditures may be justified by evaluating return on investment (ROI), which is calculated either in terms of monetary gain after the initial investment is recovered or in terms of a percentage of the initial investment.

$$\text{ROI}_{5\text{year}} = \frac{(\text{Cost avoidance} + \text{New revenue})_{5\text{year}} - \text{Investment}}{\text{Investment}} \times 100\%$$

Consider a 5-year ROI to justify acquisition of two liquid chromatography/mass spectrometry (LC/MS) systems to be used as open platforms for LDTs. The cost difference between sendout testing and in-house testing (cost avoidance), and additional revenue obtained by developing novel tests on LC/MS instrumentation, considered cumulatively over 5 years, constitute gains attributable to the initial investment. While new revenue may be difficult to predict, the cost of sendouts is known and the cost of in-house testing can be estimated in an objective manner. The initial investment includes not only the instruments but also construction costs or other costs not included in the in-house cost per reportable test. For the purpose of capital equipment justification, the instrument cost should not be included in the in-house cost per reportable calculation.

Example 6.37: *The combined cost of two LC/MS systems is $1,000,000. Construction costs are estimated at $27,000. Three tests X, Y, and Z will be developed to replace corresponding sendout tests, and in-house CPRs are given below:*

Test	Annual test volume	Reference lab cost	In-house CPR
X	*600*	*$150*	*$5.35*
Y	*700*	*$125*	*$4.65*
Z	*4200*	*$35*	*$4.65*

A novel diagnostic Test N will be developed as well. Based on a market study, $50,000 in additional annual revenue should be attainable by offering Test N. Assuming development and validation of all tests requires 1 year before savings can be realized, and calculate the 5-year ROI.

First, calculate the annual cost avoidance by finding the difference between the sendout and in-house costs for Tests *X*, *Y*, and *Z*. No savings or revenue will occur in the first year during validation, so multiply by 4 years.

Cost avoidance
= Annual volume × (Sendout cost per test—In-house CPR)
Test *X* $600 \times (\$150 - \$5.35) = \$86,790$ per year
Test *Y* $700 \times (\$125 - \$4.65) = \$84,245$ per year
Test *Z* $4200 \times (\$35 - \$4.65) = \$127,470$ per year

Cost avoidance over four years post validation =
$4 \times (\$86,790 + \$84,245 + \$127,470) = \$1,194,020$

After Test *N* is developed and validated, four years of new revenue are expected ($4 \times \$50{,}000 = \$200{,}000$). Thus the 5-year ROI calculation is

$$\text{ROI}_{5\text{year}} = \frac{(\text{Cost avoidance} + \text{New revenue})_{5\text{year}} - \text{Investment}}{\text{Investment}} \times 100\%$$

$$= \frac{(\$1,194,020 + \$200,000) - \$1,027,000}{\$1,027,000} \times 100\% = 36\%$$

References

[1] R.L. Bertholf, D.B. Endres, W. Clarke, D.G. Rhoads, C.D. Hawker, R.H. Christenson, Preanalytical variation, methods development, calculations, statistics, and evidence-based medicine, in: A.H.B. Wu (Ed.), Self-Assessment in Clinical Laboratory Science II, AACC Press, Washington, DC, 2008, pp. 1–43.

[2] J.B. Campbell, J.M. Campbell, Laboratory Mathematics: Medical And Biological Applications., fifth ed., Mosby, Maryland Heights, MO, 1997.

[3] L. Buckingham, Fundamental Laboratory Mathematics: Required Calculations for the Medical Laboratory Professional, F.A. Davis Company, Philadelphia, PA, 2014.

[4] L.J. Doucette, Mathematics for the Clinical Laboratory., third ed., Elsevier, St. Louis, MO, 2016.

[5] A. Kallner, Laboratory Statistics: Methods in Chemistry and Health Sciences, second ed, Elsevier, Cambridge, MA, 2018.

[6] Ready Reference Guides. Merck manual: professional version. <http://www.merckmanuals.com/professional/appendixes/ready-reference-guides/ready-reference-guides>, 2018 (accessed 28.09.18).

[7] Weisstein, Eric W. Logarithm. MathWorld--a Wolfram web resource. <http://mathworld.wolfram.com/Logarithm.html>, 2018 (accessed 28.09.18).

[8] R.E. Karcher, J.P. Landers, Electrophoresis, in: C.A. Burtis, E.R. Ashwood, D.E. Bruns (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, fifth ed., Elsevier, St. Louis, MO, 2012, pp. 287–305.

[9] G.J. Schwartz, A. Muñoz, M.F. Schneider, R.H. Mak, F. Kaskel, B.A. Warady, et al., New equations to estimate GFR in children with CKD, J. Am. Soc. Nephrol. 20 (3) (2009) 629–637.

[10] L.A. Inker, C.H. Schmid, H. Tighiouart, J.H. Eckfeldt, H.I. Feldman, T. Greene, et al., Estimating glomerular filtration rate from serum creatinine and cystatin C, N. Engl. J. Med. 367 (1) (2012) 20–29.

[11] J.S. Chew-Harris, P.K. Chin, C.M. Florkowski, P. George, Z. Endre, Removal of body surface area normalisation improves raw-measured glomerular filtration rate estimation by the Chronic Kidney Disease Epidemiology Collaboration equation and drug dosing in the obese, Intern. Med. J. 45 (7) (2015) 766–773.

[12] National Kidney Foundation. Glomerular filtration rate (GFR). <https://www.kidney.org/professionals/KDOQI/gfr> (accessed 28.09.18).

[13] National Institute of Diabetes and Digestive and Kidney Diseases. Glomerular filtration rate (GFR) calculators. <https://www.niddk.nih.gov/health-information/communication-programs/nkdep/laboratory-evaluation/glomerular-filtration-rate-calculators> (accessed 28.09.18).

[14] M.R. Langlois, M.J. Chapman, C. Cobbaert, S. Mora, A.T. Remaley, E. Ros, et al., Quantifying atherogenic lipoproteins: current and future challenges in the era of personalized medicine and very low concentrations of LDL cholesterol. A consensus statement from EAS and EFLM, Clin. Chem. 64 (7) (2018) 1006–1033.

[15] G.H. White, I. Farrance, Uncertainty of measurement in quantitative medical testing: a laboratory implementation guide, Clin. Biochem. Rev. 25 (4) (2004) S1–S24.

[16] CLSI, Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures; Approved Guideline–Second Edition. CLSI Document EP17-A2, Clinical and Laboratory Standards Institute, Wayne, PA, 2012.

[17] L.S. Garcia, P. Bachner, V.S. Baselski, M.R. Lewis, A.J. Linscott, D.A. Schwab, et al. (Eds.), Clinical Laboratory Management, second edition, ASM Press, Washington, DC, 2014.

Chapter 7

Spectrophotometry

Lorin M. Bachmann and W. Greg Miller
Department of Pathology, Virginia Commonwealth University, Richmond, VA, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Explain the principles of spectrophotometric measurement.
- Describe the sources of error in spectrophotometric measurement.
- Discuss the calibration of spectrophotometric methods.
- Explain the principles of related optical measurement technologies such as fluorescence, chemiluminescence, and atomic emission.

Principles of light absorption and emission

Spectrophotometry is the measurement of the absorption of electromagnetic radiation, in the wavelength range of ultraviolet (UV), visible, and near-infrared (IR) light, by an atom or a molecule to determine the quantity or identity of the substance present. In clinical chemistry, this analytical technique is used to measure the quantity of an analyte present in a body fluid sample. In most cases, an analyte of clinical interest participates in a chemical reaction resulting in the formation of a product, called a chromogen (or chromophore), which absorbs light at a specific measuring wavelength. The measuring wavelength is selected based on the wavelength, or wavelength range, at which the chromogen exhibits peak absorbance. The quantity of chromogen produced in the reaction is proportional to the quantity of analyte, and therefore the amount of light absorbed by the chromogen is proportional to analyte concentration in the clinical sample. Spectrophotometry measures the amount of light absorbed to quantitate the chromogen. A related analytical technique measures the light emitted as fluorescence. In fluorescent measurement techniques, a specific wavelength of light is absorbed by a chromogen, called a fluorophore, and the subsequent emission of a longer wavelength of light is measured. Other complimentary photometric techniques measure light emitted from atoms when stimulated by nonelectromagnetic energy sources (e.g., a gas flame or plasma) or from molecules that emit light when stimulated by a chemical reaction (chemiluminescence).

The energy state of the molecular orbitals of the organic molecules can be represented as illustrated in Fig. 7.1. At typical laboratory temperatures, the electrons are in the ground (lowest energy) state, but the energy is distributed among several vibrational and rotational energy levels. A molecule can transition to an excited-state electron orbital distribution if sufficient energy is available for the transition. For molecules with an appropriate molecular orbital electron distribution, absorption of radiation in the 200−1000-nm wavelength range can cause a transition to a stable higher energy orbital. Such a transition is illustrated in Fig. 7.1. The arrows represent electrons going from ground-state energy levels to excited-state levels. Absorption of radiation is quantized between specific electronic, vibrational, and rotational energy levels, which means that only the exact quantity of energy required for a transition to a given higher energy level will be absorbed. Because of quantized absorption, a molecule absorbs specific wavelengths corresponding to the energy required for a transition. The individual quantized transitions are not resolved at typical laboratory temperatures, because there are many transitions with similar energy because of multiple vibrational and rotational states. For this reason, it is typical to observe relatively broad chromogen absorption bands (absorbance of multiple wavelengths within a given wavelength range) in liquids at ambient temperatures.

After absorption of radiation, the excited-state molecule can dissipate the energy in several ways. One is for the excited state to undergo a photochemical reaction with another molecule to produce a new compound; this process is not discussed in this chapter. In spectrophotometric applications, thermal transfer primarily dissipates the energy through collision with other molecules in the liquid solution. Therefore no energy in the form of photons is emitted. In fluorescence applications, the excited state emits radiation at a longer wavelength

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00007-7

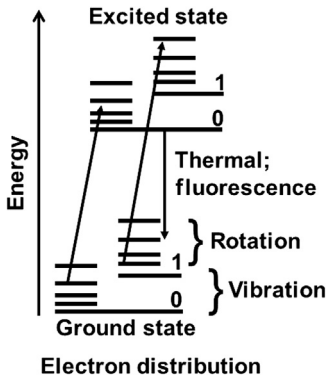


FIGURE 7.1 Diagram showing relative energy distribution of electrons among rotational and vibrational levels in a ground state and a first excited state of an orbital, and quantized energy absorption or loss (*arrows*) when an electron transitions between the states.

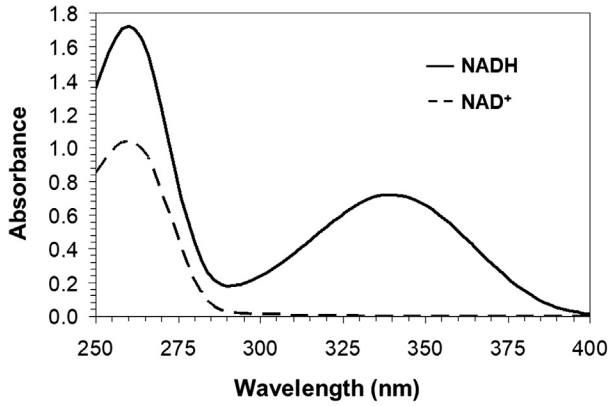


FIGURE 7.2 Absorption spectrum of the NAD^+/NADH chromogens.

(lower energy), as it relaxes to the ground state. Not shown in Fig. 7.1, and infrequently used in clinical chemistry, is the possibility for singlet excited-state electrons to undergo further electronic orbital transitions to a triplet excited state and emit radiation at a longer wavelength, called phosphorescence, which persists for a longer time interval.

Typical chromogens that are absorbed in the UV−visible−IR wavelength range are organic molecules with highly conjugated C = C, C = O, C = N, and other double bonds and/or aromatic rings that have molecular orbital electron distributions that can transition to an excited state at the energies available in these wavelength ranges. Metallic atoms can also absorb energy in this wavelength range to produce excited-state electronic orbitals.

The absorption spectrum describes chromogen absorbance vs. the wavelength range of light passing through a cuvette that contains a solution of the chromogen. Fig. 7.2 shows an absorption spectrum for the oxidized and reduced forms of nicotinamide adenine dinucleotide (NAD^+ and NADH, respectively). Many spectrophotometric assay reaction mechanisms utilize the absorptive properties of the NAD^+/NADH redox pair. In this reaction scheme, an analyte-specific chemical reaction results in product formation that is coupled to the reduction of NAD^+ to NADH or oxidation of NADH to NAD^+. As shown in Fig. 7.2, NAD^+ and NADH absorb light at 250−280 nm because of absorption by the adenine base. However, an absorbance specific for NADH also occurs at approximately 340 nm because of the orbital distribution of the electrons in the dihydronicotinamide ring of NADH. In contrast, NAD^+ exhibits minimal absorbance at 340 nm. An example of an analyte reaction with the NAD^+/NADH redox pair is the hexokinase method for measuring glucose.

Hexokinase:

$$\text{Glucose} + \text{ATP} \rightarrow \text{glucose-6-phosphate} + \text{ADP} \qquad (7.\text{i})$$

Glucose-6-phosphate dehydrogenase:

$$\text{Glucose-6-phosphate} + NAD^+ \rightarrow \text{6-phosphogluconate} + \text{NADH (absorbance at 340 nm)} \qquad (7.\text{ii})$$

During the second step of the reaction, enzymatic conversion of glucose-6-phosphate to 6-phosphogluconate is coupled to the reduction of NAD^+ to NADH, and absorbance is measured at 340 nm. The resulting increase in absorbance due to NADH formation is directly proportional to the concentration of glucose in the sample. It is important to note that consumption of a chromogen is also utilized in some spectrophotometric reactions. If the reaction mechanism results in the oxidation of NADH to NAD^+, the chromogen (NADH) is converted into a nonchromogenic product at 340 nm (NAD^+), as shown in a method for measuring lactate dehydrogenase (LDH) activity.

LDH:

$$\text{Pyruvate} + \text{NADH (absorbance at 340 nm)} \rightarrow \text{lactate} + NAD^+ \qquad (7.\text{iii})$$

In this case, absorbance decreases in proportion to pyruvate concentration conversion to lactate by LDH.

Other methods of chromogen production include dye-binding reactions, chemical conversion of an analyte to a chromogen, and enzyme-mediated conversion of an analyte to a chromogen. In the dye-binding methods, direct binding of an analyte to a precursor compound in the reagent leads to the formation of a colored complex that absorbs light at a specific wavelength. An example of a dye-binding reaction is a commonly used method for measuring total calcium.

$$Ca^{2+} + \text{arsenazo III} \rightarrow \text{Ca:arsenazo III complex (absorbance at 660 nm)} \qquad (7.\text{iv})$$

Dye-binding reactions are used for many analytes, including orthocresolphthalein complexone for calcium,

pyrogallol red and Coumassie brilliant blue for total serum protein, and bromcresol green and bromcresol purple for serum albumin.

For quantitation of analytes using spectrophotometric procedures, it is typical to select a single wavelength to measure the amount of light energy absorbed by the chromogen. The wavelength corresponding to the maximum absorbance is usually used to optimize the analytical sensitivity of the measurement. However, the influence of interfering substances is a consideration when selecting a measuring wavelength as discussed in a later section.

Principles of spectrophotometric measurement

Fig. 7.3 illustrates a spectrophotometric measurement system. Light of the desired wavelength impinges on a photodetector that converts the light energy into an electrical signal. A container, called a cuvette, is placed in the light path of the spectrophotometer. The cuvette is filled with a blank solution that, ideally, contains all reaction components except the chromogen, and the amount of light hitting the photodetector is measured (called I_0). The cuvette is then filled with a solution that contains all reaction components, including the chromogen, and the light hitting the photodetector is measured again (called I). The amount of light hitting the photodetector is reduced by the absorbance of light by the chromogen. The amount of chromogen produced by the reaction and the corresponding magnitude of absorbance are proportional to the amount of analyte present in the cuvette. A spectrophotometric measurement is the ratio of I to I_0 and is called the transmittance (T) or the percent transmittance ($\%T$) when the ratio is multiplied by 100. The formula is

$$T = I/I_0 \tag{7.1}$$

The transmittance ratio has an inverse, semilogarithmic relationship to the concentration of the chromogen. The transmittance is more commonly converted into absorbance (A), which has a linear relationship to concentration to make it more convenient to use. It should be noted that transmittance is the parameter actually being measured and that the corresponding absorbance is a calculated value. The formula is

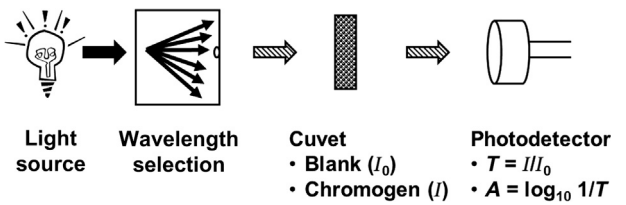


FIGURE 7.3 Diagram showing major components and configuration of a typical single-beam spectrophotometric system.

$$A = \log_{10}(1/T) = -\log_{10} T \tag{7.2}$$

Beer's law describes the linear relationship between absorbance and concentration.

$$A = (\varepsilon)\ (b)\ (C) \tag{7.3}$$

where A is absorbance, ε is molar absorptivity [L/(mol•cm)] of the chromogen, b is path length (cm) of light through the cuvette, and C is concentration (mol/L) of the chromogen in the solution in the cuvette. The molar absorptivity is a physical property of the chromogen. This property is defined as the amount of light, at a specified wavelength, that a 1 mol/L concentration of the chromogen absorbs assuming the cuvette path length is 1 cm. The molar absorptivity can be affected by the temperature, pH, and matrix of the solution containing the chromogen. Variations in temperature, pH, and matrix (e.g., ionic strength, protein content, and surfactants), as well as errors in wavelength and bandpass, can cause an error when calculating concentration using Beer's law. For these reasons, many methods correct for such variations by calibration, which will be discussed in more detail later in the chapter.

It is important to remember that a spectrophotometric measurement is a ratio of I to I_0, because it is the intensity of light hitting the photodetector that limits the measuring range for a chromogen concentration. An absorbance of 0.0 is 100%T; an absorbance of 1.0 corresponds to 10%T, which means the chromogen has absorbed 90% of the incident light energy. Similarly, an absorbance of 2.0 corresponds to 1%T, which means the chromogen has absorbed 99% of the incident light energy. Consequently, the concentration change that causes an absorbance change from 0 to 1 will use 90% of the measurement range of the spectrophotometer; doubling the concentration will cause the absorbance to change from 1 to 2 but will only use 9% of the measurement range of the spectrophotometer. The progressively smaller change in the ratio of I to I_0 as concentration increases and the corresponding increasing contribution of noise sources to the signal become a limiting factor in the ability of a spectrophotometer to measure accurately the concentration of a chromogen. For this reason, the optimal measurement range for most spectrophotometers is between 0 and 1.0A. Some instruments can make acceptable measurements at larger A values, but there will be a progressive decrease in precision and accuracy at higher values.

Configuration of spectrophotometers

Figs. 7.3 and 7.4 show two common configurations for a single-beam spectrophotometric measurement system. A wavelength selection device can be placed before the cuvette as in Fig. 7.3 or after the cuvette as in Fig. 7.4,

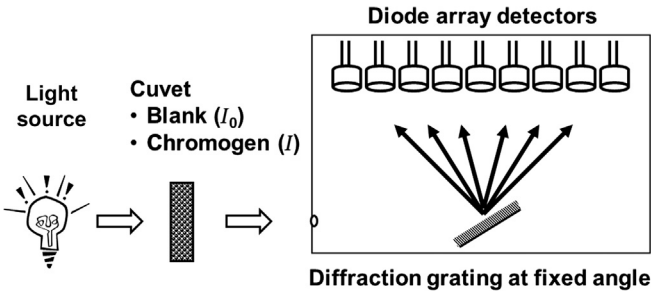


FIGURE 7.4 Diagram showing the configuration of a diode array spectrophotometric system.

depending on the design of the spectrophotometer. In Fig. 7.3, the specific wavelength of interest is isolated, passes through the cuvette, and hits the detector. This configuration can use a monochromator or an interference filter to isolate the wavelength of interest. In Fig. 7.4, the broad spectrum of wavelengths passes through the cuvette, is diffracted by a grating monochromator, and hits a series of individual diode detectors (called a diode array configuration). Each diode in the array is impinged by a very narrow band of wavelengths such that many different wavelengths can be measured simultaneously. The diode array configuration has no moving parts to go out of alignment. Diode array configurations are common in automated analyzers because of their stability and the ability to measure simultaneously multiple wavelengths. Research-grade diode array spectrophotometers can measure in wavelength increments of 1 nm. However, many clinical laboratory instruments are restricted to specific intervals of wavelengths that may limit flexibility for development of new assays.

Absorbance is measured by placing a cuvette containing the solution without chromogen in the light path measuring I_0. The solution in the cuvette is then replaced with a solution containing the chromogen and I is measured. Alternatively, the chromogen-containing solution could be in a separate cuvette with a path length identical to that of the one used for I_0 measurement. This configuration can easily be incorporated into a typical random access automated analyzer in which a single reference cuvette is used for the I_0 measurement. A series of cuvettes that contain the analyte-specific chromogens is measured to obtain I for a series of clinical samples. The ratio of I/I_0 for each cuvette/reference combination is then used to determine the absorbance for each sample. In addition to providing an I_0 value for a specific set of measurements, the reference channel can also be used to monitor the stability of the spectrophotometric system over time and to correct for optical or electronic signal drift.

Another common configuration for a spectrophotometer is called a double-beam system, in which the light path is split with mirrors and is directed alternately through each of two cuvettes, one used for the I_0 and one for the I measurements. The frequency of the light alternating between the two cuvettes is typically fast enough to permit the measurement of transmittance over a range of wavelengths to obtain an absorption spectrum.

Critical operating parameters: accuracy

Bandpass

Spectral bandpass (bandwidth) refers to the range of wavelengths that pass through the cuvette when the monochromator precedes the cuvette or can be isolated to impinge on an individual detector in a diode array configuration. The intensity of light isolated at a specific wavelength will be most intense at the nominal selected wavelength, but a range of wavelengths with progressively decreasing intensity will also be transmitted. The bandpass is defined as the range of wavelengths at the point of half-intensity of transmitted light. Consequently, a range of wavelengths is actually measured by the detector. If the molar absorptivity of the chromogen changes over the bandpass of the instrument, which is common, there will be less absorbance observed than expected for the concentration of chromogen present. This condition results in a lower than expected absorbance, because the observed absorbance is the integral of absorbance over the band of wavelengths passing through the cuvette and hitting the detector. If the conversion of absorbance to concentration is based on the molar absorptivity (e.g., in methods that measure enzyme activity), there will be an error in the concentration measured unless the molar absorptivity is corrected for the actual bandwidth of a specific spectrophotometer.

For example, the NADH chromogen shown in Fig. 7.2 exhibits a change in absorbance vs. wavelength. If a spectrophotometer with a wavelength bandpass of 1 nm was used to measure NADH, there would be little absorbance error, because 1 nm is small enough to isolate functionally the peak absorbance at which the molar absorptivity was determined. However, if a spectrophotometer with 10-nm bandpass was used, some of the wavelengths would not be measuring the peak absorbance, the integrated absorbance over the bandpass would be less than that expected, and the concentration calculated from the molar

absorptivity would be erroneously low. For NADH, a 10-nm bandpass would produce an absorbance that is approximately 1% lower than the correct value and a 20-nm bandpass would produce an absorbance that is approximately 3% lower than the correct value. Bandpass has a greater influence for chromogens with a narrower absorption spectrum and a smaller influence for those with a broader absorption spectrum. A bandpass of <8 nm has been recommended when measuring NADH in enzyme assays that determine activity on the basis of molar absorptivity [1].

When a relatively wide bandpass instrument is used, it is necessary to correct for the condition by determining a functional molar absorptivity value specific for that instrument. Instrument manufacturers have various names (e.g., filter factor) for such instrument-specific molar absorptivity values. When interference filters are used, the instrument-specific molar absorptivity values must be verified periodically, because the devices are subject to deterioration over time. Alternatively, an instrument-specific calibration can be established using a calibrator with known concentration (or activity) to establish the relationship between absorbance and concentration (or activity) for that instrument.

Stray light

Stray light is the light at wavelengths other than the measuring wavelength that reaches the detector and is not absorbed by the chromogen. Stray light causes the measured absorbance to be lower than it would be in the absence of the stray light. Consider the equation for transmittance when a stray light component (I_S) is present.

$$T = \frac{I + I_S}{I_0 + I_S} \tag{7.4}$$

The intensity of light hitting the detector has the stray light added to it. I_S is usually an insignificant fraction of I_0, but as I gets small due to large absorbance by the chromogen, I_S becomes a significant fraction of I, which causes T to be larger (and A to be lower) than what would be expected based on the concentration of chromogen. Consequently, the relationship between absorbance and concentration becomes nonlinear, as stray light becomes a greater percentage of the total light hitting the detector. For example, if stray light is 0.1% of the total hitting the detector, the absorbance will have a 1% error at approximately 1.5A, and if stray light is 0.5%, the absorbance will have a 1% error at approximately 0.6A.

Stray light can be caused by light leaks in the instrument, degradation of the source lamp, and, rarely, fluorescence or phosphorescence from endogenous substances or drug metabolites in the sample. However, the most common source of stray light is undesired wavelengths originating from the monochromator or wavelength isolation device because of optical defects. Contemporary holographic grating monochromators have reduced stray light to insignificant levels for many applications, but this source of error can be an issue when interference filters are used. Some spectrophotometric systems have incorporated automated multiwavelength stray light correction algorithms, which are effective for some types of optical defects.

Stray light can be measured as residual absorbance when measuring a solution or a filter that absorbs $>4A$ (0.01 %T) at the wavelength of interest but passes most other wavelengths. A signal observed in these conditions is due to stray light reaching the detector. Note that only potential stray light at wavelengths that are not blocked by the filter or solution used to detect stray light will be identified by this procedure. Two or more test filters or solutions are typically needed to cover a range of wavelengths. Therefore stray light checks should be performed as part of routine instrument performance assessment. Use of a pulsed-frequency xenon lamp as the spectrophotometer source lamp is a common technique used to reduce the effects of light leaks. The light from the xenon lamp is pulsed at a frequency much faster (e.g., 1000 Hz) than that of the overhead fluorescent lights (60 Hz) and is synchronized with the photodetector. This approach reduces the error contribution of minor amounts of environmental light contamination.

Wavelength accuracy

The wavelength selected by the monochromator or other device must be verified to be correct as indicated. If the wavelength is not as indicated, the molar absorptivity used to calculate concentration will be incorrect for the actual wavelength in use, corrections for interfering substances may not be accurate, or the analytical sensitivity of the measurement may not be as expected. Wavelength accuracy can be verified by scanning and identifying the peak absorbances for known substances such as holmium oxide in solution [National Institute of Standards and Technology Standard Reference Material (NIST SRM) 2034] or in glass, or by removing wavelength-isolating filters from an instrument and scanning them in a calibrated spectrophotometer. Some spectrophotometers have a provision to use sharp line emissions from deuterium or mercury lamps used as UV light sources. Verifying wavelength accuracy is the responsibility of the clinical chemist for a stand-alone spectrophotometer, but is the responsibility of the manufacturer for a spectrophotometer built into a diagnostic instrument.

Absorbance accuracy

The absorbance measured by the spectrophotometer must be verified to be correct whenever results are based on

molar absorptivity. Verification of absorbance accuracy can be accomplished by measuring a set of certified absorbance standards. Liquid solutions with a known concentration of chromogen with a known molar absorptivity can be used. A set of liquid solutions for the range 302−678 nm (SRM 931h) and solid potassium dichromate (SRM 136f) can be obtained from NIST to create liquid solutions for use at wavelengths 235−350 nm. Another approach (useful for automated analyzers for which it is difficult to add solutions directly into cuvettes) is to generate quantitatively NADH, for which the molar absorptivity is known at 340 nm, from pure glucose (NIST SRM 917c) using the hexokinase reaction. Verifying absorbance accuracy is the responsibility of the clinical chemist for a stand-alone spectrophotometer but is the responsibility of the manufacturer for a spectrophotometer built into a diagnostic instrument.

Chromogen limitations

The measured chromogen is usually generated by a chemical reaction that converts the amount of analyte present in a biological sample into a proportional amount of the chromogen. The proportionality of the chemical conversion can be affected by defects in the reagents or by an inadequate amount of one or more reagents that may limit the completeness of the chemical reaction(s). The chromogen, at higher concentrations, may form complexes or adducts with itself or with other components of the reaction that can change the molar absorptivity and thus the ability to transform accurately the absorbance to the concentration of the analyte.

Calibration of spectrophotometric measurements

Calibration establishes the mathematical relationship between the analytical response and the chromogen concentration in the solution in the cuvette. The concentration of the analyte in the biological sample is typically directly proportional to the analytical response of the chromogen. Spectrophotometric measurements are based on the molar absorptivity of the chromogen or on a calibration relationship established by a calibrator that compensates for the unique conditions of an instrument. It is a common practice to calibrate a spectrophotometric assay using the calibration solutions of known analyte concentration, which are assayed in the same manner as the clinical samples, in which case the calculated concentration of an analyte in an unknown sample is not reliant on knowing the molar absorptivity. However, spectrophotometric assays for enzyme activity typically use the molar absorptivity of a chromogen to calculate the activity as the rate of substrate consumption. Errors related to the molar absorptivity are important to control when Beer's law is used to calculate enzyme activity or analyte concentration.

Molar absorptivity

When concentration is calculated from the molar absorptivity, it is imperative that the pipette, absorbance, wavelength and temperature accuracy, bandpass, and stray light have been verified. Each of these variables can affect the chemical reaction(s) and the observed molar absorptivity of the chromogen in the cuvette. Any variable that can influence the matrix of the solution in the cuvette (e.g., pH, ionic strength, surfactants, and temperature) must be controlled to ensure the chromogen is in the same matrix conditions as when the molar absorptivity was determined. Note that it is acceptable and recommended to establish the molar absorptivity of the chromogen for the unique conditions of a specific method and spectrophotometric system. Note that some automated analyzers may not permit the user to make changes to molar absorptivity settings. A measurement system-specific molar absorptivity can be established by preparing a series of solutions containing various amounts of pure chromogen dissolved in the reaction buffer for the method. The molar absorptivity is calculated as the absorbance per mole per liter of chromogen in the cuvette for a range of concentrations.

Calibration relationship

A calibration relationship can be established by measuring a series of concentrations of calibrators (also called standard solutions) and establishing the mathematical relationship between the absorbance and concentration of analyte in the calibrator solutions. The calibrator solutions are measured in the same manner as the clinical samples and must have a matrix that is appropriate for the clinical samples that will be quantitated using the calibration relationship. The term "standard curve" or "calibration curve" is used to refer to a calibration relationship even when the relationship is a straight line. For most spectrophotometric assays, the relationship between absorbance and concentration is linear, although the relationship between transmittance (the actual measured quantity) and concentration is semilogarithmic. If the absorbance relationship is not linear, several concentrations of the calibrator are required to describe accurately the calibration relationship over the analytical measurement range.

Establishing a calibration relationship eliminates or substantially reduces the influences of biases in pipette, wavelength, and absorbance accuracy, bandpass, and stray light, because these variables are compensated for by having the calibrators measured at the same spectrophotometric conditions as the unknown samples. These variables remain important for the measurement and must be at

settings appropriate for the reactions involved, but nominal differences from specifications can be tolerated as long as they remain constant for calibrators and unknown samples. The need for temperature accuracy is eliminated as far as its effect on the molar absorptivity of the chromogen but the role of temperature on the chemical reaction that generates the chromogen may still be important.

Precision of the various steps in the assay remains an important variable to control. The matrix of the calibrator solutions must be appropriate for the clinical samples to be assayed to ensure there is no bias introduced by a difference in the extent of the chemical reactions involved in generating the chromogen or in the final matrix of the solution in the cuvette that might differently affect the molar absorptivity of the chromogen, depending on whether it originated from the calibrator or from the clinical sample. The matrix should also be taken into consideration when diluting a sample. The selected diluent should possess matrix characteristics similar to the sample to avoid measurement biases due to matrix differences.

One of the critical factors affecting measurement accuracy is the target value assignment of the calibrator materials. The calibrator value assignment should be traceable to a higher order reference measurement procedure, if one is available, or to a commutable certified reference material when available and no reference measurement procedure exists. Reference measurement procedures and reference materials can be found in the database of the Joint Committee on Traceability in Laboratory Medicine (JCTLM) [2]. Many of the reference materials listed by JCTLM are intended for use as trueness controls with reference measurement procedures and are not commutable with clinical samples when used with routine clinical laboratory methods. Use of noncommutable reference materials as calibrators for routine clinical laboratory methods can cause miscalibration and erroneous results for clinical samples [3,4]. See Chapter 17, Harmonization of results among laboratories, for additional information on calibration traceability.

Critical operating parameters: precision

The precision of a spectrophotometric measurement varies with design of the instrument system. At low absorbance (high $\%T$), the precision is limited by mechanical, optical, and electronic stabilities. At high absorbance (low $\%T$), the precision is also limited by stray light and detector background electronic noise, as these sources of error become a larger fraction of the measured signal. Maximum precision is generally obtained when the absorbance is $0.1-1.0A$. Contemporary automated spectrophotometers typically make multiple independent absorbance measurements and average the results to reduce the effects of imprecision. Precision improves proportional to the square root of the number of independent measurements (e.g., the mean of nine measurements reduces imprecision by a factor of 3; 100 measurements reduces imprecision by a factor of 10).

The precision of a spectrophotometric measurement can be validated in several ways depending on the accessibility of the cuvette. A dye can be measured in replicate in a single cuvette or in multiple cuvettes. When multiple cuvettes are used, minor differences in path length among cuvettes will be included in the imprecision value. For automated systems, it is frequently adequate to measure the combined effects of pipette and absorbance variation on the precision using a dye (which removes any influence of a chemical reaction) or with a chemical reaction to generate a chromogen that may also include reagent formulation variability and temperature effects.

Interferences

The absorbance of a substance other than the chromogen derived from the analyte of interest will cause interference in the measurement, unless the signal from the interfering substance is removed from the total absorbance. Absorbance from the reagents used in a reaction can be measured before adding the sample and subtracted from the total absorbance of the reagent plus chromogen. This process is called a reagent blank correction. Clinical samples such as serum or urine frequently contain substances that may absorb light at the wavelength used to measure the analyte-specific chromogen and cause spectral interference. The effects of spectral interferences can be reduced in several ways, including the use of a sample blank, use of a kinetic reaction-rate method, and correction using wavelengths at which the interfering substances absorb.

Measurement of a sample blank can correct for spectral interferences in spectrophotometric end-point methods. In an end-point reaction (Fig. 7.5), the absorbance value is obtained at a time point after the reaction has proceeded to completion (when all of the analyte has reacted with the reagents to produce a chromogen). To correct for spectral interferences, a "sample blank" absorbance reading at the measuring wavelength is taken after mixing the sample with all of the reagents, except those that cause chromogen production. The reagent that triggers production of chromogen is added, and the reaction is allowed to proceed to completion. The total absorbance is measured at the completion of the reaction. The absorbance of the sample blank, corrected for the volume of the added second reagent, is then subtracted from the total absorbance to yield the corrected chromogen-specific absorbance. This method of correction is subject to error if the absorbance of the spectral interferant is of such a large magnitude that the chromogen-specific contribution

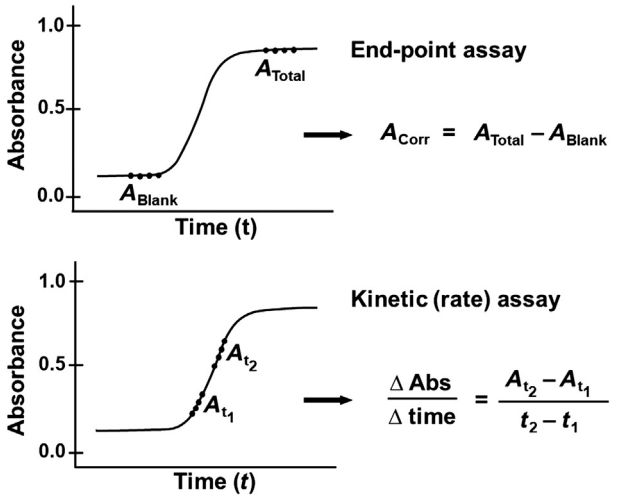


FIGURE 7.5 Spectrophotometric end-point and rate reactions. The first inflection on each graph represents the addition of reagents required for chromogen production. *Note*: Equations are simplified examples; automated instruments typically include polychromatic corrections for interfering substances, sample volume corrections, and averaging of many measurements to improve precision.

to the total absorbance is small. It also is important to note that sample or reagent blanks can only correct for interfering absorbances that remain constant during the reaction. The magnitude of the correction will be erroneous if an absorbing interferant reacts with any component of the reaction mixture.

Another procedure to correct for spectral interferences is by utilization of a kinetic (rate) reaction (Fig. 7.5). Rather than calculating the analyte concentration on the basis of total absorbance at the end of a reaction, the absorbance is measured at multiple time points during the reaction. The change in absorbance vs. time is then used to calculate analyte concentration. Measurement of the change in absorbance over time corrects for the absorbance of interfering substances that do not change during the course of the reaction. As is the case with sample blanking, this correction method will not work if the interfering substance causes a large absorbance, such that the change in absorbance due to the analyte-specific chromogen cannot be accurately measured.

An important limitation to the preceding correction techniques occurs when the absorbance from an interfering substance changes during the time of the reaction. Examples of dynamic absorbance interference include clearing of turbidity due to dispersion of triglyceride by surfactants and reaction of an interfering chromogen such as bilirubin in peroxide-coupled reactions as shown in the following example:

Glucose oxidase:

$$\text{Glucose} + O_2 \rightarrow H_2O_2 + \text{products} \qquad (7.v)$$

$$H_2O_2 + \text{dye} \rightarrow \text{chromogen (analyte-specific absorbance)} \qquad (7.vi)$$

$$H_2O_2 + \text{bilirubin} \rightarrow \text{nonchromogen (decrease in total absorbance)} \qquad (7.vii)$$

Note that the interference from bilirubin has two mechanisms. First, some H_2O_2 that would otherwise be available to react with the dye to produce chromogen instead reacts with bilirubin and is not converted into the analyte-specific chromogen, thus resulting in a falsely low absorbance. Second, bilirubin itself can absorb light and may contribute to the total absorbance. The interfering absorbance from bilirubin decreases in magnitude over the time course of the reaction. As the bilirubin reacts with H_2O_2, the absorbance from bilirubin no longer contributes to the total absorbance. Dynamic interferences must be controlled by modifying the reaction system to remove the interfering substance independently from the reaction that generates the analyte-specific chromogen. For example, inactivating agents such as sodium dodecyl sulfate (SDS) or potassium ferricyanide can be added to the reagent to reduce the interference of bilirubin for reaction methods that use peroxide to produce the chromogen such as those based on the Trinder reaction. In some instances, the interfering substance exhibits different reaction kinetics than the reaction of the analyte. In these cases, the time points at which absorbance readings are obtained can be modified to reduce the effect of the interference.

Another limitation in the spectrophotometric methods, as in all methods, is caused by the nonspecificity of the reaction for the analyte of interest. In this case, interfering substances react to produce the same chromogen that is produced from the analyte of interest. For example, acetoacetic acid can react with alkaline picrate used in some creatinine methods to produce a chromogen that cannot be discriminated from that produced from creatinine. Method nonspecificity must be corrected by modifying the method to include additional separation steps to remove interfering substances or to change the chemical mechanism to be more specific for the analyte. When interferences exist in a method, they must be documented as limitations in method specificity.

Bichromatic and polychromatic measurements

Measuring absorbance at a second wavelength at which no change in absorbance occurs during the reaction can be used as a reference point to correct the spectrophotometric system for drift in the measurement signal. Two or

more wavelengths can also be used to correct for some types of absorbance interferences. A simple case occurs when an interfering substance absorbs light at the same wavelength as the analyte-specific chromogen (primary wavelength) but also absorbs light at a second wavelength at which the analyte-specific chromogen does not absorb. The ratio of molar absorptivities of the interfering substance at the two wavelengths can be used to predict the absorbance due to the interfering substance at the primary wavelength from its absorbance at the second wavelength. This value can then be subtracted from the total absorbance at the primary wavelength to obtain the net absorbance due to the analyte-specific chromogen. This correction approach has been referred to as the Allen correction.

A more complex situation occurs when the analyte-specific chromogen and an interfering substance both absorb light at all wavelengths available for measurement. This situation is shown in Fig. 7.6. When the molar absorptivities of the analyte-specific chromogen and the interfering substance(s) are different and are known at two wavelengths, two equations with two unknowns can be written to describe the total absorbance at each of the two wavelengths. This correction method has been referred to as bichromatic correction. A is the absorbance; λ_1 and λ_2 are the two respective wavelengths; ε_1 and ε_2 are the molar absorptivities of substances 1 and 2, respectively, at each of the two wavelengths; and C_1 and C_2 are the concentrations of substances 1 and 2, respectively.

At wavelength 1:

$$A_{(\lambda 1)} = \left(\varepsilon_{1(\lambda 1)} \times C_1\right) + \left(\varepsilon_{2(\lambda 1)} \times C_2\right) \tag{7.5}$$

At wavelength 2:

$$A_{(\lambda 2)} = \left(\varepsilon_{1(\lambda 2)} \times C_1\right) + \left(\varepsilon_{2(\lambda 2)} \times C_2\right) \tag{7.6}$$

All arguments in the two equations are known or can be measured except C_1 and C_2. Thus the two equations can be simultaneously solved for C_1 and C_2, one of which is the concentration of the analyte of interest.

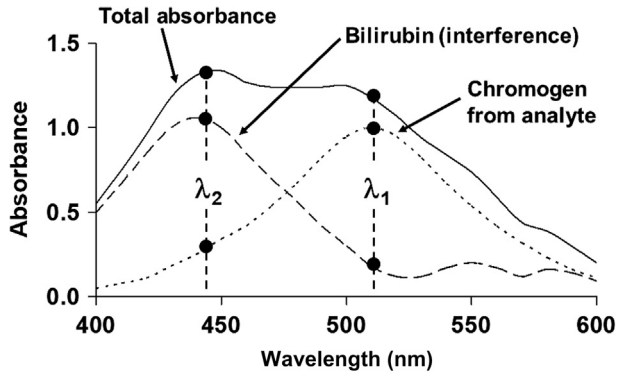


FIGURE 7.6 Absorption spectra illustrating bichromatic measurement to correct for an interfering substance that has absorption at all wavelengths available to measure a chromogen.

Note that more than one interfering substance can be corrected using a bichromatic (two wavelengths) technique as long as the substances have similar molar absorptivities at the two wavelengths used. For example, bilirubin and hemoglobin have very similar absorbance spectra at several wavelengths and can frequently be adequately corrected using bichromatic measurements. However, interfering substances that have different molar absorptivity ratios at the two wavelengths cannot be corrected using bichromatic measurements. For example, turbidity from triglyceride-rich lipoproteins has a very different absorption spectrum than bilirubin and hemoglobin; thus a bichromatic technique cannot simultaneously correct for each of these sources of spectrophotometric interference. Three (trichromatic) or more wavelengths can be used to correct for more complex interferences or to measure simultaneously multiple analytes, such as in cooximetry (COOX) determination of hemoglobin species. Imprecision of a final result will increase as more individual absorbance measurements are used and must be accommodated in the total error budget of the measurement.

Other applications of spectrophotometric or light emission measurements

Cooximetry

COOX is the measurement of blood total hemoglobin and the relative percentages of hemoglobin derivatives including oxyhemoglobin (O2Hb), deoxyhemoglobin (HHb), carboxyhemoglobin (COHb), and methemoglobin (MetHb). COOX instrumentation may be integrated with routine blood gas analyzers or function as stand-alone analyzers. O2Hb, HHb, COHb, and MetHb each exhibit characteristic spectrophotometric absorbance profiles over the wavelength range of 422−695 nm. The absorbances at the measuring wavelengths are obtained for venous or arterial whole blood samples, and a series of algorithms based on the absorbance spectrum for each molecule is used to identify and quantify each hemoglobin derivative. Total hemoglobin is quantified using conductivity or the sum of the spectrophotometric measurements of O2Hb, HHb, COHb, and MetHb. Due to similar absorbance spectra, sulfhemoglobinemia from medications containing sulfonamides can cause interference with MetHb measurements and fetal hemoglobin can cause interference with COHb measurements. Therefore COOX instruments employ additional algorithms for qualitative spectrophotometric detection of sulfhemoglobin and fetal hemoglobin to alert the user of elevated concentrations of these derivatives. Although different combinations of instrument

optics and measuring wavelengths are used to reduce the effects of interferences, high concentrations of bilirubin, intralipid, or methylene blue can also cause errors in COOX measurements. More information on the use and interpretation of COOX measurements can be found in Chapter 35, Contemporary practice in clinical chemistry: blood gas and critical care testing.

Reflectance spectrophotometry

Reflectance refers to a configuration in which the incident light impinges on the chromogen contained in a thin coating on a surface, and the light is reflected from the rear surface of the coating support to make a second pass through the chromogen matrix. Examples include dry slide chemistry analyzers and urinalysis dipstick readers. In these applications, the path length is small and measurement precision is influenced by the uniformity of the coating containing reagent(s) and chromogen. Reflectance is characterized by increased influence of scatter at each reagent layer surface that has a change in the index of refraction. Consequently, a diffuse reflectance measurement typically has a nonlinear relationship with concentration and appropriate calibration algorithms are used.

Densitometry

Densitometry measures absorbance by passing light through a solid support material on which the chromogen is contained by absorption or adsorption to the solid (or gel) layer. Thin-layer chromatography and electrophoresis are typical examples. Accuracy and precision are limited by variations in path length due to inconsistent thickness of the medium, scatter from components in the solid medium, and exceed the limited dynamic range of the detector caused by too high of a concentration of the chromogen in the medium.

Turbidimetry and nephelometry

Turbidimetry and nephelometry are the techniques to measure concentration on the basis of light scatter, as illustrated in Fig. 7.7. Conventional spectrophotometers measure light transmitted at 180 degrees from (or in line with) the incident light source. Turbidity is measured in this configuration as absorbance, because scattered light does not reach the photodetector and causes a decrease in I_0 (increase in absorbance), just as does absorbance by a chromogen. This mechanism is the reason for which turbidity causes interference in spectrophotometric measurements. Nephelometry measures the intensity of light scattered at an angle away from the incident light path. Nephelometry is not a measurement of the ratio of light intensity in the absence and presence of scattering molecules but is a measure of absolute intensity of light caused by scattering molecules. Nephelometry is capable of measuring orders of magnitude lower concentrations of scattering molecules than turbidity. Nephelometers frequently use laser light sources, because the increased intensity can produce increased intensity of scattered light and lower detection limits. In addition, optics and optical filters are used to focus the scattered light and reduce the contribution of stray light to improve the analytical sensitivity of the measurement.

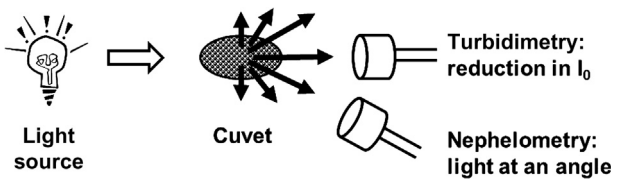


FIGURE 7.7 Configuration used to measure light scatter by turbidimetry or by nephelometry.

Light is scattered by molecules large enough that their hydrated diameter is of similar magnitude to the wavelength of the incident light [5]. Light scatter is typically used to quantitate antibody complexes of protein analytes, because these complexes are large enough to scatter light. Light scatter is a complex process that depends on the size and index of refraction of the scattering molecular complex. Light scatter is typically greater at shorter wavelengths, with the intensity being approximately proportional to the inverse of the wavelength to the fourth power ($1/\lambda 4$). Light scatter is not symmetrical; the intensity of scatter at a given angle from incident varies with the size of the complex. As the scattering complex gets larger, a greater proportion of light is scattered at angles >90 degrees from the incident (called forward scatter). Nephelometers typically measure scatter at forward angles from the incident to improve sensitivity for immune complexes and similar sized molecular aggregates. It is common to control the size of immune complexes by the use of surfactants and polymers (e.g., polyethylene glycol) to optimize conditions for quantitation of complexes of different sizes that may be produced by analytes at different concentrations. Measurement of light scatter can be used to measure amylase and lipase hydrolysis of starch or triglycerides, respectively, and for assays based on the aggregation of latex particles coated with various substances, including antibodies. Flow cytometry measures scatter of light by cells (typically larger than the molecular aggregates) at small forward angles to determine the size and concentration of the cells.

The response of turbidimetry and nephelometry is not linear with the concentration of analyte because of the variable size of scattering complexes formed at different analyte concentrations. When the scattering complexes are formed by the classical antibody–antigen reaction,

the complex size is susceptible to antigen-excess conditions. Antigen excess may cause falsely low results at very high concentrations of analyte (called the "high-dose hook effect" or "prozone effect"). To detect a high-dose hook effect, the absorbance of the undiluted sample is compared with the absorbance of a sample that has been diluted. If the absorbance of the diluted sample is greater than the undiluted sample, a prozone effect is present. Most modern spectrophotometers and nephelometers automatically analyze the kinetics of antigen−antibody complex formation to increase the probability of prozone effect detection and alert the user.

Nephelometry and turbidimetry are sensitive to interferences from light scatter due to triglyceride-rich lipoproteins (lipemia). The effects of interfering lipoproteins may be partially reduced by ultracentrifugation and removal of the lipid particles. Because turbidimetric methods are based on the measurement of transmitted light, they are also sensitive to the absorbance of colored substances such as hemoglobin and bilirubin in the sample. Antibodies are often utilized as detection reagents, and therefore the presence of cross-reacting substances or interfering antibodies (e.g., rheumatoid factor or heterophile antibodies) can be a source of measurement error.

Atomic absorption

Atomic absorption is a special application in which ground-state atoms of metals absorb light at very specific wavelengths corresponding to the energy needed to cause electronic transitions in their electron orbitals. Fig. 7.8 shows the configuration of an atomic absorption spectrophotometer. The atoms are created by heating the sample in an acetylene gas flame or a cylindrical graphite tube furnace to a temperature that burns off most of the organic matrix and produces ground-state atoms of the metal to be measured. The "cuvette" is the flame or the center of the graphite tube, and the atoms are in the gas state. The population of atoms is relatively short-lived in the light path, and rapid measurements are necessary. The thermal energy must be carefully controlled and ionization of metals must be avoided maintaining the atoms in the ground state.

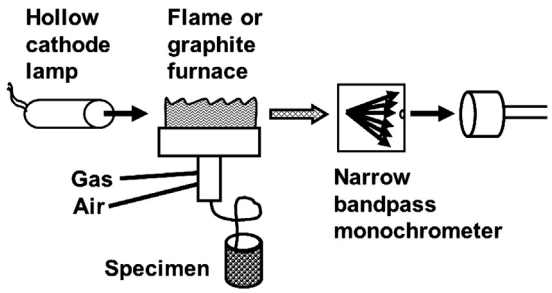


FIGURE 7.8 Diagram showing major components and configuration of an atomic absorption spectrophotometer.

Atomic electron transitions have very narrow bandwidths that require specialized light sources with very narrow bandpass (a few angstroms or tenths of nanometers). Hollow cathode lamps have an electrode coated with the metal to be measured and use an electronic discharge to generate an emission wavelength from the metal that is at the correct wavelength to be absorbed by the ground-state atoms in the cuvette. A high intensity of stray light is produced by incandescence of sample materials in the flame or furnace that necessitates placing the monochromator after the cuvette in the light path to the detector. A very narrow bandpass monochromator is used to isolate the wavelength of interest for the metal from the stray light. Because of the stray light, linearity of atomic absorption is frequently limited to a small range of absorbance (e.g., 0−0.3). Background absorbance and light scatter from sample components can also interfere and can be corrected by specialized bichromatic (Zeeman effect) or polychromatic (deuterium lamp or Hieftje) techniques. Atomic absorption can be used for accurate assay of many metals in body fluid and tissue samples, because most sample interferences can be eliminated by correct choice of instrument and thermal conditions.

Atomic emission

Atomic emission is a photometric technique in which thermal energy is used similarly to atomic absorption, except that, in this technique, the thermal energy is adequate to create excited-state atoms, which then emit light as they relax back to the ground state. Atomic emission offers the possibility to measure simultaneously several different atomic species and can have lower limits of quantitation than atomic absorption. The emitted light is at wavelengths specific to the metals present. Flame photometry using propane gas provides the reference measurement procedures for sodium, potassium, and lithium in clinical samples. An inductively coupled argon ion plasma (ICP) can be used to create excited-state atoms for many metals. The ICP is much hotter than propane or acetylene flames and thus produces more complete combustion of potential interfering organic substances. The hotter plasma also produces more emission lines from metals and requires a more sophisticated optical system to isolate wavelengths of interest. ICP methods are used for toxic and trace metal measurements from body fluid and tissue samples.

Fluorescence

Fluorescence is the light emitted from an excited-state molecular orbital as it relaxes to the ground state by

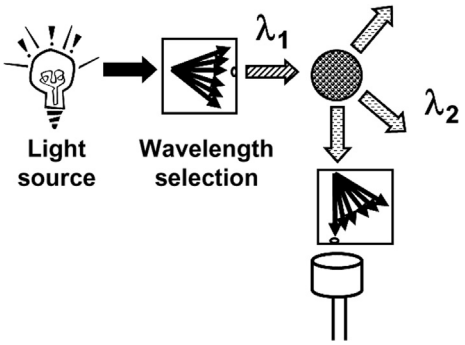


FIGURE 7.9 Diagram showing major components and typical configuration to measure fluorescence.

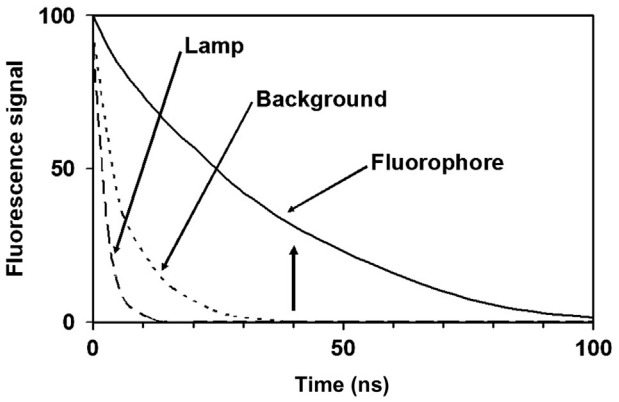


FIGURE 7.10 Time course of fluorescence emission after a pulse of excitation light.

emitting a photon of light. Fluorescent molecules, called fluorophores, are characterized by the extended conjugation of double bonds with aromatic resonance that allows the lifetime of the excited state to persist long enough for relaxation to occur by the emission of light. The excited state is produced by the absorption of light that has greater energy (shorter wavelength) than that of the emitted light. The emitted light is always at a longer wavelength (lower energy), because the excited-state electrons rapidly lose thermal energy to reach a minimal energy excited state before the emission of light to relax to the ground state. The difference in wavelength between the excitation and emission wavelengths is called the Stokes shift. Fig. 7.9 shows a fluorescence measurement configuration. Fluorescent light is emitted with equal intensity in all directions, which allows for various optical configurations. Fluorescence intensity is measured at an angle and using a wavelength selection device that excludes the excitation light. The fluorescence intensity is directly proportional to the concentration of a fluorophore in the cuvette. The quantum yield is the fraction of absorbed energy that is emitted by the fluorophore and is a measure of the efficiency of molecular fluorescence.

Fluorescence typically provides a 100-fold to 1000-fold increase in concentration sensitivity compared with absorbance measurements. For maximum sensitivity, the detection system must be able to measure the fluorescence emission without any background signal from the excitation light (i.e., measure the signal against a zero background). Potential interferences are from substances in the sample that can quench the excited-state molecule before fluorescent light is emitted or that can absorb the emitted light in the cuvette before it reaches the detector (called an inner filter effect).

Fluorescence lifetime

Fluorescence emission typically persists for 10−100 ns after excitation to the excited state, depending on the fluorophore, until all of the excited-state molecules have returned to the ground state. When longer lifetime fluorophores are used, this characteristic can be exploited to eliminate interferences from substances that contribute background fluorescence emission and to improve the limit of quantitation. Fig. 7.10 illustrates a time-resolved fluorescence measurement when a relatively long-lifetime fluorophore is used. A pulse of light is used for excitation; when the pulse ends, the background fluorescence decays to 0 in 30 ns, but the fluorescence from the long-lifetime fluorophore persists and can be measured against a zero-signal background at 40 ns.

Fluorescence polarization (depolarization)

If a fluorescent molecule absorbs polarized light (i.e., all electromagnetic waves are oriented in the same direction) to produce the excited state, then the fluorescence emission will have the same polarization as long as the fluorophore does not rotate during the lifetime of the excited state. If the fluorophore rotates before the fluorescence emission takes place, the emission will be depolarized proportional to the degree of rotation before the light emission. This property has been the basis for fluorescence polarization (more correctly depolarization) immunoassay for low-molecular-weight haptens. In this application, a fluorescent analog of the analyte and the analyte from the sample are mixed and compete for analyte-binding sites on a fixed number of antibodies. If analyte is absent, the fluorescent analog binds to the antibodies and rotates in solution at the rate of the antibodies. Because an antibody is a large protein, it rotates relatively slowly compared with the low-molecular-weight fluorescent analyte analog, such that there is little change in position during the lifetime of the excited state. Consequently, the fluorescence polarization of the emitted light is not changed from that of the excitation light. If analyte is present, it binds to some of the antibody molecules, which leaves some of the fluorescent analog unbound to antibodies and free in solution. Because the

fluorescent analog is small, it rotates relatively quickly, such that its orientation changes during the lifetime of the excited state before fluorescent light is emitted. Consequently, the fluorescence polarization of the emitted light is randomized or depolarized. The amount of depolarization or the ratio of polarized to depolarized signal is the basis to quantitate the analyte. Historically, the fluorescence polarization approach was used in several immunoassay instruments but is not used in current practice.

Chemiluminescence

Chemiluminescence is related to fluorescence in that light emitted from a molecular orbital excited state is measured and is proportional to the concentration of the chemiluminescent substance. In this technique, the excited state is produced by a chemical reaction rather than by absorption of light, as in fluorescence excitation. The excited chemiluminescent compound emits light as it relaxes to a ground electronic energy state. Many precursor compounds used in chemiluminescent reactions (e.g., acridinium ester, luminol, and luciferin analogs) are activated to emit light upon oxidation. In the presence of an oxidizing agent, metal ions or enzymes (e.g., horseradish peroxidase or alkaline phosphatase) mediate oxidation of the precursor compounds. Chemiluminescent reactions often have a wide dynamic range and can have greater sensitivity to low concentrations than absorption or fluorescence, because the measurements are made against a zero-signal background.

The most common applications for chemiluminescent measurement are immunoassay methods, in which the chemiluminescent compound or enzyme mediator functions as part of the detection label. Light emission is detected by a luminometer, in which emitted photons are detected to produce a measurement signal. The analytical sensitivity of chemiluminescent reactions is improved by the inclusion of a second molecule (sometimes called an intensifier agent) that can amplify the intensity of emitted light by a nonradiative energy transfer mechanism or stabilize the excited state to optimize light emission. Common sources of error in clinical practice include the development of light leaks in the luminometer and insufficient removal of unbound chemiluminescent label because of defects in the wash system. Like fluorescence, potential interferences include molecules that can quench the excited state before light emission and molecules that can absorb the emitted light in the cuvette before it reaches the detector.

References

[1] Guidelines for photometric instruments for measuring enzyme reaction rates. Clin. Chem.23 (1977) 2160–2162.

[2] Joint Commission on Traceability in Laboratory Medicine Database. <https://www.bipm.org/jctlm/> (accessed 13.11.18).

[3] W.G. Miller, G.L. Myers, R. Rej, Why commutability matters, Clin. Chem. 52 (2006) 553–554.

[4] H.W. Vesper, W.G. Miller, G.L. Myers, Reference materials and commutability, Clin. Biochem. Rev. 28 (2007) 139–147.

[5] G. Sittampalam, G.S. Wilson, Theory of light scattering measurements as applied to immunoprecipitation reactions, Anal. Chem. 56 (1984) 2176–2180.

Suggested reading

J.R. Lakowicz, Principles of Fluorescence Spectroscopy, Springer, New York, 2006. 954pp.

H.A. Strobel, W.R. Heineman, Chemical Instrumentation: A Systematic Approach, third ed., Wiley, New York, 1989. 1210 pp.

H.H. Willard, L.L. Merritt, J.A. Dean, F.A. Settle, Instrumental Methods of Analysis, seventh ed., Wadsworth Publishing, Belmont, CA, 1988. 895 pp.

Self-assessment questions

(More than one answer may be correct for each question.)

1. What is actually measured to determine the absorbance of a chromogen?
   - **a.** light energy hitting a detector
   - **b.** decrease in light intensity hitting a detector
   - **c.** increase in light intensity hitting a detector
   - **d.** ratio of light intensity hitting a detector between the chromogen and the blank
2. Which of the following reaction types can be used to produce chromogens that absorb light?
   - **a.** redox reactions
   - **b.** chemiluminescent reactions
   - **c.** dye-binding reactions
   - **d.** enzyme-mediated reactions
3. What can cause the relationship between absorbance and concentration to be nonlinear?
   - **a.** stray light
   - **b.** too large of a bandpass
   - **c.** chromogen complexation
   - **d.** wavelength calibration error
4. Why is 0.1−1.0 the most precise measurement range for absorbance?
   - **a.** Effects of stray light are minimized.
   - **b.** The proportion of light intensity change with concentration decreases at larger *A* values.
   - **c.** The noise of the detector becomes dominant at low *A* values.
   - **d.** The noise of the detector becomes dominant at high *A* values.
5. What can cause the apparent molar absorptivity to be different from theoretical?
   - **a.** too large of a bandpass
   - **b.** stray light
   - **c.** inner filter effect
   - **d.** miscalibrated wavelength
6. What type of calibration can make molar absorptivity unimportant?
   - **a.** calibration relationship using calibrator solutions
   - **b.** correct wavelength calibration
   - **c.** bracketing with standards
   - **d.** use of a filter factor
7. What sources of bias or error can be corrected for by calibration?
   - **a.** wavelength inaccuracy
   - **b.** autopipettor imprecision
   - **c.** absorbance inaccuracy
   - **d.** spectrophotometric interferences
8. How can the absorbance from an interfering substance be corrected?
   - **a.** sample blank before final reagent addition
   - **b.** kinetic measurement
   - **c.** bichromatic measurement for a sample with hemoglobin and turbidity
   - **d.** rapid absorbance measurement
9. What type of interference cannot be corrected with a bichromatic measurement?
   - **a.** bilirubin
   - **b.** turbidity
   - **c.** bilirubin and turbidity in the same sample
   - **d.** nonchromogenic drugs
10. What is the relationship between turbidimetric and nephelometric measurements?
    - **a.** Nephelometry is the inverse of turbidimetry.
    - **b.** Turbidimetry is more sensitive.
    - **c.** Nephelometry can be measured with a conventional spectrophotometer.
    - **d.** Both measure light scatter.
11. How can the presence of a prozone effect in an immunoturbidimetric assay be detected?
    - **a.** measuring the absorbance after ultracentrifugation
    - **b.** measuring the absorbance after concentrating sample
    - **c.** measuring the absorbance after diluting the sample
    - **d.** measuring the absorbance after treating the sample with SDS
12. How does atomic absorption spectrophotometry differ from solution-phase measurements?
    - **a.** The chromogen has broad absorbance bands in atomic absorption.
    - **b.** Stray light is more significant in atomic absorption.
    - **c.** The chromogen is short-lived in atomic absorption.
    - **d.** The flame or furnace in atomic absorption produces a more stable chromogen.
13. How does fluorescence polarization immunoassay discriminate between free and bound fluorescent analogs?
    - **a.** The bound fluorescent analog emits polarized light in all directions.
    - **b.** The bound fluorescent analog emits polarized light with the same orientation as that of the light absorbed.
    - **c.** Free fluorescent analog rotates during the lifetime of fluorescence and depolarizes the emitted light.
    - **d.** Fluorescent-labeled analog and native analyte compete for antibody-binding sites.
14. Which of the following are sources of error in chemiluminescent assays?
    - **a.** insufficient wash
    - **b.** presence of substances that quench emission
    - **c.** presence of light leaks
    - **d.** lamp deterioration

Answers

1. d
2. a, c, d
3. a, c
4. a, b, d
5. a, b, d
6. a, c
7. a, c
8. a, b, d
9. c
10. d
11. c
12. b, c
13. b, c
14. a, b, c

Chapter 8

Chromatography and electrophoresis

David S. Hage
Department of Chemistry, University of Nebraska, Lincoln, NE, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Define "chromatography" and describe common ways of categorizing different types of this method.
- Discuss the terms that are used to describe retention, band-broadening, and peak separations in chromatography.
- Explain the significance of the van Deemter equation and the general factors that affect the resolution of a separation in chromatography.
- Discuss the general concepts of gas chromatography and liquid chromatography, including the types of mobile phases, elution methods, supports, and stationary phases that are used in these techniques.
- Describe the design and applications of common detectors and detection systems that are used in gas chromatography and liquid chromatography.
- Define "electrophoresis" and describe general formats for this method.
- List the various factors that can affect the migration and band-broadening of chemicals in electrophoresis.
- Discuss the general concepts of gel electrophoresis and capillary electrophoresis, including the types of supports, application methods, and formats that are used in these techniques.
- Describe common detectors or detection systems that are used in gel electrophoresis and capillary electrophoresis.

Introduction to chromatography

General terms and concepts

The complexity of clinical samples such as blood, serum, and urine typically necessitates some separation prior to analysis. Chromatography is a method that is frequently employed for this purpose. Chromatography is a separation technique in which the components of a sample are separated based on how they distribute between two chemical or physical phases, a stationary phase and a more mobile phase that is allowed to travel through the separation system. This method is illustrated in Fig. 8.1, in which a sample is applied to the top of a tube, known as the "column." This column contains a fixed surface or layer, known as the "stationary phase," that is held in place by a solid support and that can interact with the various components of the sample. A second phase (called the "mobile phase") is used to apply samples to the column and pass their components through to the other end. Those substances that have the weakest interactions with the stationary phase will travel through more quickly than those with strong interactions, resulting in a separation of these chemicals [1–4].

There are many different types of chromatography. The main way of categorizing these techniques is based on their mobile phase. If a gas is used as the mobile phase, the method is called gas chromatography (GC); if the mobile phase is a liquid, the technique is known as liquid chromatography (LC). Both GC and LC can be divided into further subcategories based on their separation mechanism, which is determined by the type of stationary phase in the column. For instance, the use of underivatized solid particles in GC or LC as the stationary phase produces the methods of gas–solid chromatography (GSC) and liquid–solid chromatography (LSC).

Another way chromatographic methods can be grouped is according to the type of chromatographic bed or support that is present. A method in which a column contains the support and stationary phase is known as column chromatography. If the column is packed with support particles, this method is called packed bed chromatography. If the stationary phase is instead placed directly onto the interior wall of the column, the technique is known as open-tubular chromatography. It is also possible to have the support and stationary phase present on a flat plane, giving the technique of "planar chromatography." Specific examples of planar chromatography include paper chromatography, which uses paper as a support, and thin-layer chromatography, which generally uses a stationary phase coated on a glass or plastic sheet.

One way that chromatography can be utilized is for the isolation of a given chemical prior to its analysis by a

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00008-9

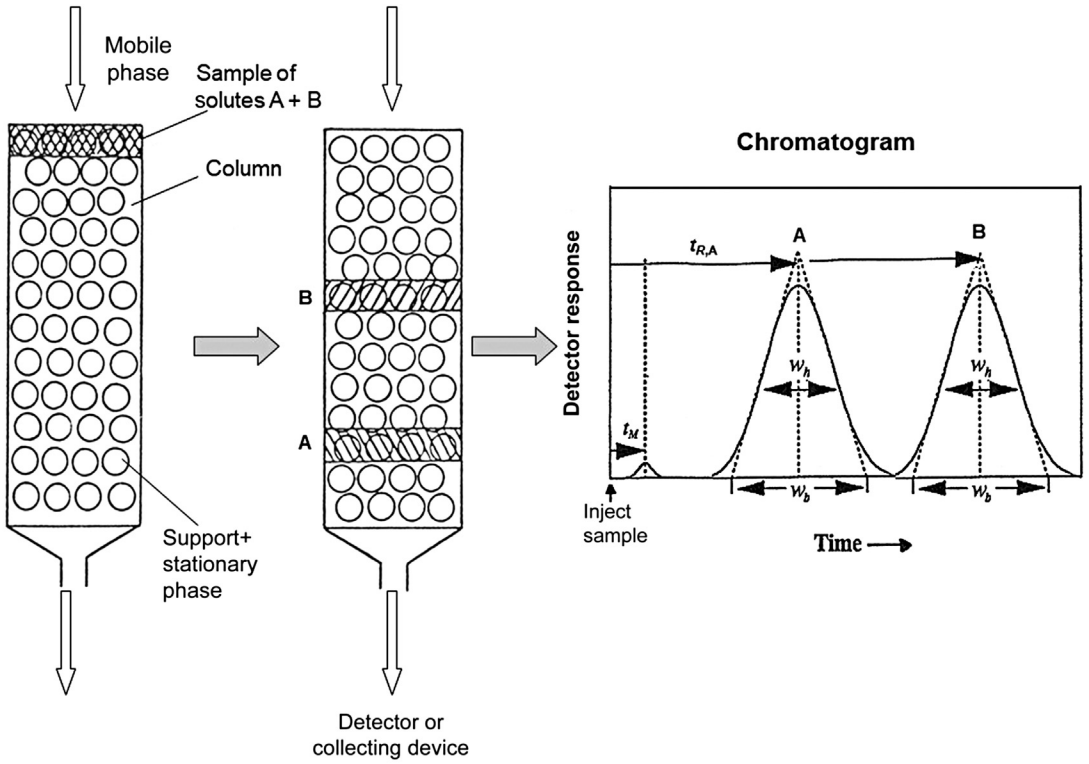


FIGURE 8.1 (A) General scheme for the separation of chemicals by chromatography and (B) a typical chromatogram for such a separation.

different method. However, chromatography can also be used directly for the identification and measurement of sample components if it is used along with a detector that can monitor these chemicals as they pass through the chromatographic system. The result is usually plotted by making a graph of the response measured by this detector as a function of the time or volume of mobile phase needed for the elution of each injected chemical. This plot, as shown in Fig. 8.1, is known as a "chromatogram."

Theory of chromatography

In chromatography, there is always a minimum amount of time required for any substance to pass through the system. This is known as the column "void time" (t_M). If a compound is bound (or "retained") by the stationary phase, it will travel more slowly through the column and exit at some later time. The average time required for this is known as that compound's "retention time" (t_R) (see Fig. 8.1). The length of this retention time is determined by the eluting substance's structure, as well as the type and composition of stationary and mobile phases being used in the chromatographic system. Elution volume can also be used to describe the movement of substances in chromatography. The average volume of mobile phase it takes to move a compound through the column is called that compound's "retention volume" (V_R). Similarly, the volume of mobile phase it takes to elute a totally nonretained substance is known as the column "void volume" (V_M).

Another way the retention of substances in chromatography can be described is by using the retention factor (k). The retention factor can be calculated by using one of the following relationships.

$$k = \frac{(t_R - t_M)}{t_M} \quad \text{or} \quad k = \frac{(V_R - V_M)}{V_M} \tag{8.1}$$

The retention factor is a more fundamental measure of retention than t_R and V_R, because it is related directly to the amount of stationary phase in the column and the extent to which solutes are interacting with this phase. As an example, Eq. (8.2) is another way of describing the retention factor for a chromatographic system in which a solute is partitioning between the mobile phase and the stationary phase.

$$k = K_D\left(\frac{V_S}{V_M}\right) \tag{8.2}$$

In this equation, (V_S/V_M) (i.e., the phase ratio) is the relative volume of stationary phase versus the volume of mobile phase in the column, and K_D is the distribution equilibrium constant for the solute in these two phases. Eq. (8.2) indicates that the degree of retention in chromatography will be related to both the degree of that

compound's interactions with the stationary phase (K_D) and the relative amount of stationary phase in the column (V_S/V_M).

As chemicals travel through a chromatographic system, the width of the region that contains each compound (i.e., the compound's peak or band) gradually becomes broader. This process is known as "band-broadening" and occurs even for substances that have little or no binding to the stationary phase. One way of describing band-broadening in chromatography is to measure the width of a peak at its baseline level (w_b) or its width at half-height (w_h) (Fig. 8.1). However, these values are highly dependent on the retention time or retention volume used for measurement. To compare the band-broadening for substances with different retention times, the "number of theoretical plates" or "plate number" (N) is utilized. The value of N can be calculated from w_b or w_h by one of the expressions given in Eq. (8.3) for a peak with a Gaussian shape.

$$N = 16\left(\frac{t_R}{w_b}\right)^2 \quad \text{or} \quad N = 5.545\left(\frac{t_R}{w_h}\right)^2 \tag{8.3}$$

A large value for N makes it easier for a chromatographic system to separate compounds. One way this parameter can be viewed is by having the value of N represent the number of times a substance will equilibrate between the mobile phase and stationary phase as it travels through the column. If a compound is able to undergo more of these equilibration steps, it becomes possible for the system to distinguish between compounds with smaller differences in their retention.

Although N is useful for describing band-broadening in columns, it does have the disadvantage of being dependent on the column length. This problem can be overcome by using a related value known as the "height equivalent of a theoretical plate" or "plate height" (H). The value of H is calculated from N by using Eq. (8.4), where L is the length of column.

$$H = \frac{L}{N} \tag{8.4}$$

The value of H gives the length of column that corresponds to one theoretical plate, or one equilibration step of the analyte with the stationary phase. Because a large number of theoretical plates are desirable to provide a small degree of band-broadening, a small plate height is also desirable.

Plate heights are used not only to describe band-broadening and column efficiency, but also to evaluate how a particular experimental factor impacts band-broadening. One well-known relationship developed for this purpose is the van Deemter equation, which shows how a change in linear velocity (u, which is directly proportional to flow-rate) alters the measured plate height for a column (H).

$$H = A + \frac{B}{u} + Cu \tag{8.5}$$

In this equation, the terms A, B, and C are constants that represent the contributions to band-broadening due to several processes that occur as a solute passes through the column. These processes are eddy diffusion (A), longitudinal diffusion (B), and stagnant mobile phase mass transfer plus stationary phase mass transfer (C), where the effect of each process on H has a different dependence on the flow-rate. Eddy diffusion is produced by the presence of the large number of flow paths around support particles, with each path having a slightly different length. Longitudinal diffusion refers to the broadening of a compound's peak due to the diffusion of solutes along the length of the column. Stagnant mobile phase mass transfer is related to the rate of movement of solutes as they go from the outside of the support to the region in the support's pores or near the support's surface. Stationary phase mass transfer is related to the rate at which an analyte transfers between the stationary phase and the mobile phase.

A plot of the van Deemter equation gives a "U"-shaped curve in which there is a distinct linear velocity where H has its lowest optimum value. This type of graph is helpful in finding the flow-rate or linear velocity that gives the best possible efficiency for a column. Another way a van Deemter plot can be used is to compare the plate heights of columns that contain different supports or that are prepared by different processes. This information is useful in the development of new columns or in comparing existing columns.

Evaluating and optimizing chromatographic separations

The final success of any chromatographic separation will depend on how well the peaks of interest are separated. There are a number of ways to evaluate the extent of this separation. One way is to use the separation factor (α, also known as the "selectivity"). The separation factor is a measure of the relative difference in retention of two solutes as they pass through a column. The value of α is calculated as shown below, where k_1 is the retention factor for the solute exiting first from the column and k_2 is the retention factor for the second.

$$\alpha = \frac{k_2}{k_1} \tag{8.6}$$

The separation factor becomes larger as the relative difference in retention increases between two peaks. This

feature makes α useful in describing the effectiveness of a separation in chromatography. The separation factor can also be used to indicate whether it is feasible to resolve two compounds by a given column, in which a value for α greater than one is needed for any separation to occur.

Another approach for describing the separation of two peaks is to use the peak resolution (R_s). The value of R_s for two adjacent peaks can be calculated through the following formula.

$$R_s = \frac{t_{R2} - t_{R1}}{(w_{b2} + w_{b1})/2} \tag{8.7}$$

In this relationship, t_{R1} and w_{b1} are the retention time and baseline width (both in units of time) for the first eluting peak, while t_{R2} and w_{b2} are the retention time and baseline width of the second peak. An important advantage of using peak resolution instead of the separation factor to describe the extent of a separation is that R_s considers both the difference in retention between two compounds (as represented by $t_{R2} - t_{R1}$) and the degree of band-broadening (as represented by w_{b1} and w_{b2}), while α only considers the difference in retention for these substances.

The lowest possible value for R_s is zero, which occurs when two peaks have exactly the same degree of retention and are not separated. A value for R_s greater than zero represents some degree of separation between peaks, with the extent of this separation becoming larger as R_s increases. Ideally, it is desirable to have no significant overlap between two neighboring peaks. This situation usually occurs when R_s is greater than 1.5, a condition said to represent "baseline resolution." In some separations, peak resolution values between 1.0 and 1.5 are also adequate. The latter often occurs if the peaks are about the same size and can be measured by using peak heights, which are less affected by overlap than peak areas.

Because the resolution between two peaks is a measure of both the difference in compound retention and band-broadening, any items that affect retention or peak width will also affect R_s. This relationship is demonstrated in Eq. (8.8),

$$R_S = \frac{\sqrt{N}}{4}\frac{(\alpha - 1)}{\alpha}\frac{k_2}{(1 + k_2)} \tag{8.8}$$

where k_2 is the retention factor for the second peak, α is the separation factor between the first and second peaks, and N is the number of theoretical plates for the column. Eq. (8.8) is called the "resolution equation" of chromatography. This equation shows that the degree of a separation in chromatography will be affected by three factors: (1) the extent of band-broadening in the column (N); (2) the overall degree of peak retention (k); and (3) the selectivity of the column's stationary phase in binding to one compound versus another (α).

Some ways the band-broadening of a chromatographic system can be lowered to improve resolution include the use of a longer column, operation at a flow-rate producing a smaller plate height, or the use of a support with better mass transfer properties (e.g., smaller diameter support particles). An improvement in resolution by increasing the degree of retention generally involves changing the mobile phase or stationary phase. Altering the selectivity of the chromatographic system is usually performed last and is the most difficult of the options for improving resolution, because this requires more complete knowledge of the types of interactions that are taking place between the injected compounds and the stationary phase.

Gas chromatography

General terms and concepts

One of the most common types of chromatography is the method of GC [2–7]. As stated earlier, GC is a type of chromatography in which the mobile phase is a gas. Fig. 8.2

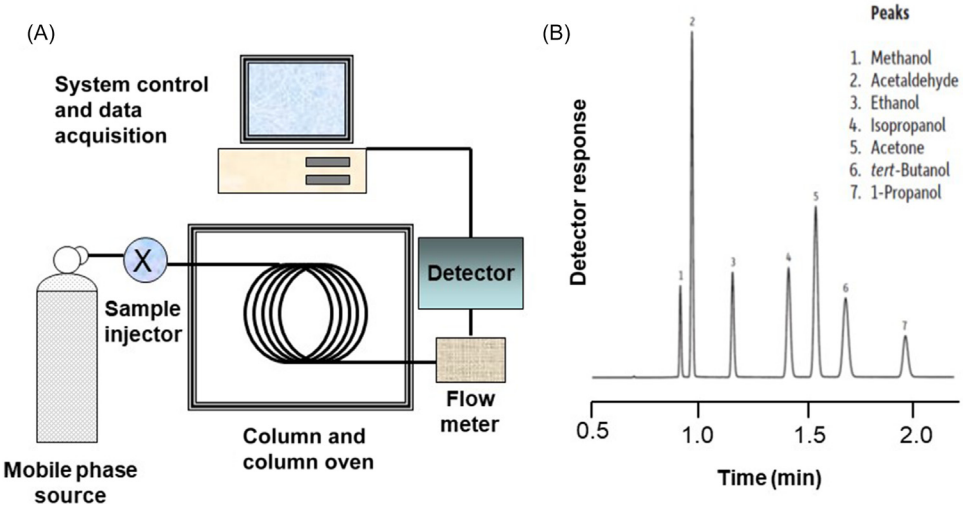


FIGURE 8.2 (A) A system for performing gas chromatography and (B) a typical separation performed by gas chromatography, using the separation of components in a blood alcohol standard as an example. *The chromatogram in (B) is adapted with permission from Restek.*

shows a typical GC system, which is known as a "gas chromatograph." The first major component of this system is the gas source that supplies the mobile phase. This source is usually a gas cylinder equipped with pressure regulators to deliver the mobile phase at a controlled rate. The second part of the gas chromatograph is the injection system, which usually consists of a heated loop or port into which the sample is placed and converted into a gaseous form. The next part of the system is the column, which contains the stationary phase and the support material. It is here that sample components are separated. This column is usually held in an enclosed area known as the "column oven" that keeps the temperature of the column and its surroundings at a well-defined level. The last part of the GC system is the detector (along with its recording device), which monitors the sample components as they leave the column.

One limitation of GC is that all analytes to be examined by this method must either be naturally volatile or capable of forming a volatile derivative. This requirement is needed so the analyte can be injected and passed through the GC system using a gas as the mobile phase. The volatility of a chemical will depend on such things as the temperature of the compound's surroundings, the size of the compound, and the types of functional groups in the compound's structure. As a rule, chemicals with molecular masses of below 600 g/mol and boiling points of below 500°C−550°C at a pressure of 1 atmosphere are usually suitable for GC. Along with good volatility, an analyte must also have good thermal stability, so that it does not decompose at the high temperatures often present in GC systems.

A common approach for increasing the volatility and thermal stability of analytes for GC is to use derivatization. This approach usually involves replacing one or more polar groups on an analyte (such as alcohol or amine groups) with less polar groups that have only weak interactions with their surrounding molecules. The result is a compound that is more volatile and easier to place into the gas phase. The same changes tend to make chemicals more thermally stable. The reaction between an alcohol-containing compound and the reagent trimethylchlorosilane is one example of derivatization. This particular reaction results in a product known as a "trimethylsilyl (TMS)-derivative" in which the hydrogen on the alcohol group is replaced with a bulkier and less polar TMS group.

Like any chromatographic method, the retention of a compound in GC will be determined by how much time this substance spends in the mobile phase versus the stationary phase. Due to the low density of gases, compounds passing through a GC column generally have no significant interactions with the mobile phase. As a result, the volatility of these chemicals is the main item that causes them to stay in the mobile phase during a GC separation. The most volatile compounds in a sample will tend to spend the most time in the mobile phase and elute most quickly from a GC column. Temperature also plays an important role in GC, because the volatility of a chemical will increase as the temperature of its surroundings increases. This is the reason why GC systems contain a column oven to maintain and control temperature during the separation process. Using a lower column temperature will produce longer retention times, because a lower temperature will cause the injected compounds to be less volatile and spend less time in the mobile phase. Increasing the temperature produces the opposite effect, with analytes becoming more volatile and passing through the column more quickly as they spend more time in the mobile phase.

Gas chromatography mobile phases and elution methods

One benefit of using a gas as the mobile phase in GC is it provides a chromatographic system with low band-broadening and narrow peaks. These narrow peaks make it easy to measure small quantities of analytes and allow GC to separate a large number of compounds in a single run. There are several reasons for the low band-broadening of GC systems. For instance, the low density of gases makes it possible for compounds in these gases to quickly move about by diffusion, which reduces many band-broadening processes. Another useful feature of gases is their low viscosity. A low viscosity for the carrier gas leads to small back pressures across GC columns, allowing the use of long columns for separations. The use of a longer column, in turn, provides a greater number of theoretical plates and better resolution.

Given that the main purpose of the mobile phase in GC is to simply move solutes through the column, the mobile phase in this technique is often referred to as the "carrier gas." Examples of common carrier gases used in GC are hydrogen, helium, nitrogen, and argon. All of these gases are relatively inexpensive, easy to obtain, inert (with the exception of hydrogen), and safe to use. These gases are usually provided by a standard gas cylinder, but sometimes they are supplied by a gas generator connected to the GC system. The carrier gas should always have high purity to avoid contamination or damage to the column and the GC system. Impurities like water, oxygen, organic substances, and particulate matter can be removed by passing the carrier gas through a series of traps and filters before it enters the column. The carrier gas source should also be equipped with regulators for pressure and flow-rate control. In some cases, it is necessary to use special devices to maintain a constant flow-rate as the temperature or pressure of the system is varied.

If a single temperature is used throughout a GC separation, this approach is called an "isothermal method." Similarly, if a constant pressure is being used, the technique is known as an "isobaric method." These elution methods work well if a sample is relatively simple or has only a few known compounds. However, these methods may not be feasible for complex samples that contain chemicals with a wide range of volatilities and/or interactions with the stationary phase. Such a situation causes some analytes to elute too quickly and others too slowly from the chromatographic system. This problem can be addressed by varying the elution conditions over time, giving an approach known as "gradient elution."

The most common way of performing gradient elution in GC is to vary the temperature of the column over time; this method is called "temperature programming." This elution method begins with a relatively low column temperature, which allows the most volatile compounds in an injected sample to better interact with the column for improved separation. Next, there is a step known as the "temperature ramp," during which the column temperature is gradually increased over time. It is during this step that analytes with intermediate volatilities are separated and eluted from the column. At the end of this ramp, the temperature is sometimes held at a high value for a short period of time to ensure that there are no low volatility substances remaining on the column from one sample injection to the next. There is then a cooling-down period in which the column is returned back to its initial temperature prior to the next sample injection.

Another type of gradient elution in GC is flow programming. In this method, the flow-rate of the carrier gas is changed during the separation. Because a change in pressure of the carrier gas is used to control its flow-rate, flow programming is sometimes referred to as "pressure programming." This method begins with a relatively slow flow-rate, which allows weakly retained components to have more time to contact the stationary phase and become better resolved. The pressure and the flow-rate are then increased to elute strongly retained compounds from the column. At the end of the elution program, the pressure and the flow-rate are adjusted back to their initial levels, and the next sample is injected. The result is a method that, like temperature programming, gives better resolution and shorter analysis times than can be obtained under constant chromatographic conditions.

Gas chromatography supports and stationary phases

GC columns can be placed into one of two major categories based on the type of support they employ: packed columns and open-tubular columns. A packed column is made up of a glass or metal tube that is usually 1−2-m long and a few millimeters in diameter; this tube is filled with small support particles that act as an adsorbent or are coated with the desired stationary phase. Packed GC columns are useful when a large amount of a sample must be separated, such as in preparative applications. This type of column can also be employed in analytical applications, but it typically has greater band-broadening than columns based on open-tubular supports. As a result, packed columns tend to be used only when a limited number of compounds are to be separated.

An open-tubular column (or "capillary" column) is used in most analytical applications of GC, including those used in clinical testing. This type of column generally ranges from 10 to 100 m in length, has an inner diameter of 0.1−0.75 mm, and possesses a stationary phase that is coated on or attached to its interior surface. A polymer such as polyimide is coated on the outside of the column to give it better strength and flexibility for handling and storage. The low band-broadening of this column allows it to provide better resolution, lower limits of detection, and/or faster rates of separation than a packed column. Three types of stationary phases are used with GC columns: (1) solid adsorbents; (2) liquids coated on solids; and (3) chemical phases bonded to a support. If a solid adsorbent is used as the stationary phase in GC, the resulting method is referred to as GSC. This technique involves the use of the same material as both the support and the stationary phase, with retention being based on the adsorption of analytes to the support's surface. An example of a support that is often used in GSC is a molecular sieve, which contains pores with well-defined sizes and binding regions. Molecular sieves are useful in retaining such things as small hydrocarbons and gases like hydrogen, oxygen, carbon monoxide, and nitrogen. Several other supports can also be used in GSC. These include organic polymers such as porous polystyrene and inorganic substances such as silica or alumina. The extent to which an analyte will bind to these materials will be determined by the support's total surface area, the size of pores in the support, and the functional groups located on the support's surface.

Another group of GC columns are those in which a chemical coating or layer is placed onto the support and used as the stationary phase. This format gives a technique known as gas−liquid chromatography (GLC), which is the most common type of GC. As shown in Table 8.1, there are many liquids used as stationary phases in this method. A common feature of all these liquids is they have high boiling points and low volatilities, which allows them to stay within the column even when high temperatures are used for sample injection and elution. Most of these stationary phases are based on polysiloxanes, which consist of a backbone of silicon and

TABLE 8.1 Typical stationary phases for gas–liquid chromatography.

Chemical name	Relative polarity[a]
100% Dimethylpolysiloxane	16 (Nonpolar)
5% Phenyl–95% methylpolysiloxane	33
14% Cyanopropylphenyl–86% methylpolysiloxane	67
50% Phenyl–50% methylpolysiloxane	119
50% Trifluoropropyl–50% methylpolysiloxane	146
50% Cyanopropylmethyl–50% phenylmethylpolysiloxane	228
Polyethylene glycol	322 (Polar)

[a]*These data are from Ewing [4]. The relative polarity is based on the McReynolds constant for benzene (X′) as a measure of stationary phase polarity.*

oxygen atoms attached in long strings of Si–O–Si bonds. The remaining two bonds on each silicon atom are bound to side groups that can have a variety of structures, ranging from nonpolar methyl groups to polar cyanopropyl groups. By altering the amount and type of these groups, it is possible to produce polysiloxanes with a variety of polarities.

One difficulty with using a liquid stationary phase is that the elevated temperatures often employed in GC will eventually cause this liquid to decompose or vaporize and leave the column over time. This process is known as "column bleeding" and changes the binding characteristics of the column. One approach for minimizing column bleeding is to use a stationary phase that is covalently attached to the support. The resulting material is known as a "bonded phase." One way a bonded phase can be produced is by reacting groups on the stationary phase with groups on the surface of the support. An alternative route is to cross-link the stationary phase to give a larger and more thermally stable structure. Either approach gives a column with less bleeding and that can be used at higher temperatures than columns that contain liquid stationary phases. It is for this reason that bonded phases are preferred in most clinical applications of GC.

Gas chromatography detectors

There are many detectors available for GC (see Table 8.2). These detectors can be divided into two categories: general detectors and selective detectors. The thermal conductivity detector (TCD) is a general detector that can be used for both organic and inorganic compounds. It produces a signal by measuring the thermal conductivity of the carrier gas/analyte mixture leaving the GC column. Hydrogen and helium are the two carrier gases that work best with a TCD, because they have the greatest difference in thermal conductivities from most analytes. The main advantage of a TCD is its ability to respond to any compound, provided that this substance is different from the carrier gas and is present in a sufficient quantity for monitoring. A disadvantage of a TCD is it will respond to impurities in the carrier gas, to stationary phase bleeding from the column, or to air leaking into the GC system. The TCD is also highly sensitive to changes in flow-rate, such as might occur during temperature programming or flow programming. Another disadvantage of a TCD is it has a relatively poor lower limit of detection compared with other common GC detectors.

The flame ionization detector (FID) is another general GC detector. An FID detects organic compounds by measuring their ability to produce ions when they are burned in a flame. The flame in the FID is usually formed by burning the eluting compounds in a mixture of hydrogen and air. Positively charged ions produced by the combustion of organic compounds in the flame are collected by an electrode and produce a current that allows these compounds to be detected. An advantage of using an FID for the GC analysis of organic compounds is it gives little or no signal for many inorganic compounds, including carrier gases (e.g., He, Ar, and N_2) and common carrier gas contaminants (O_2, CO_2, and H_2O). The FID is also less sensitive to flow-rate changes than the TCD, which makes it easier to use with temperature programming and flow programming. One disadvantage of the FID is it is a destructive detector, which breaks down analytes during the process of their measurement, preventing this detector from being connected directly to other devices or techniques for compound analysis.

A nitrogen–phosphorus detector (NPD) is a selective detector used in GC for nitrogen- or phosphorus-containing compounds. The NPD is similar to an FID in that both are based on the measurement of ions produced from eluting compounds. The main difference from an FID is that, instead of a flame, an NPD uses thermal heating at or above a rubidium surface to form ions. This process is particularly efficient for nitrogen- or phosphorus-containing compounds, which makes the NPD selective for such chemicals. The greatest strength of the NPD is its good selectivity and low limits of detection for nitrogen- and phosphorus-containing compounds. Like an FID, the NPD does not detect many common carrier gases or impurities. It is necessary when using an NPD to periodically change the heated material, because it will slowly degrade over time.

The electron capture detector (ECD) is another selective detector used in GC. This device detects compounds

TABLE 8.2 Common detectors for gas chromatography.

Detector name	Compounds detected	Detection limit[a]
Thermal conductivity detector	General: all compounds	1 ng
Flame ionization detector	All organic compounds	1 pg
Nitrogen–phosphorus detector	N- and P-containing compounds	0.01–0.1 pg
Electron capture detector	Chemicals with electronegative groups	1–100 fg
Mass spectrometry	General: full-scan mode	0.1–1 ng
	Selective: SIM mode	1–10 pg

SIM, Selected ion monitoring.
[a]*This information is from Poole and Poole [2], Ewing [4], and manufacturers of these detectors.*

that have electronegative atoms or groups in their structure, such as halogen atoms (I, Br, Cl, and F) and nitro groups (NO_2). An ECD detects compounds based on the capture of electrons by such atoms or groups. These electrons are usually produced by a radioactive source, such as ^{3}H or ^{63}Ni, and in the presence of argon or nitrogen as the carrier gas. Although an ECD is selective and has good limits of detection, it does require a radioactive source that must be changed from time to time (now commonly done by a specially trained technician from the manufacturer of this device).

Another common detector used in GC is a mass spectrometer. This combination is known as gas chromatography/mass spectrometry (GC/MS). This is a powerful tool for both measuring and identifying a compound. For instance, this approach can provide information on a chemical's molecular mass (by using its molecular ion) or its structure (through the use of its fragment ions). Common ionization methods used in GC/MS are electron impact ionization and chemical ionization. The resulting ions are separated according to their mass and/or charge by using a device such as a quadrupole mass analyzer or ion trap. One way of viewing this information is to make a plot of the number of ions measured at each elution time, giving a mass chromatogram. Another way this information can be viewed is by looking at the mass spectrum for all ions that are produced for analytes eluting at a particular retention time.

A big advantage of GC/MS is it can be used to either universally or selectively detect compounds as they leave the column. When used as a universal detector, the total number of measured ions is used to plot the mass chromatogram. This method involves using the mass spectrometer to quickly scan through a wide range of mass-to-charge ratios while collecting information on each ion that occurs within this range. This detection format is also known as the "full-scan mode" of GC/MS. This mode is useful when the goal is to look for a broad range of compounds in a single analysis (e.g., during drug screening) or when a mass spectrum is desired to determine the identity of an unknown compound (e.g., determining the drug taken by a patient suffering from an overdose). A more selective mode for performing GC/MS is selected ion monitoring (SIM). In this approach, only a few ions characteristic of the compounds of interest are examined. This mode is employed when low detection limits are desired and when it is known in advance what compounds are to be analyzed.

Methods for sample injection and pretreatment in gas chromatography

Because GC requires that analytes be in the gas phase, gaseous analytes and samples are natural candidates for this technique. If the gaseous analytes are present at moderate-to-high concentrations, direct sampling and injection onto a GC system is possible. Direct injection can be performed by passing the sample through a gas-tight valve. Alternatively, a gas-tight syringe can be used to inject a known volume of gas into the GC system. For trace components in gases, it is often necessary to collect and concentrate these analytes for separation and measurement by GC. One way this collection and concentration can be accomplished is by passing a large volume of samples through a solid-phase extraction cartridge, a cold trap, or a liquid in which the analytes will dissolve.

Liquids are the type of sample most frequently encountered in GC. When using packed columns or wide-bore open-tubular columns, it is possible to use direct injection to place this liquid into the GC system. This direct injection involves the use of a calibrated microsyringe to apply the desired volume of liquid to the system through a gas-tight septum and into a chamber where it is vaporized for analysis. However, the volumes of liquid delivered by most microsyringes are too large for the

open-tubular columns used in many clinical applications of GC. This problem can be overcome by using split injection, in which only a small portion of the vaporized sample (0.01%−10%) is allowed to go into the column.

It is quite common for some pretreatment to be required before a liquid sample can be analyzed by GC. An example is the derivatization of a fatty acid or amino acid into a more volatile or thermally stable form prior to an analysis by GC. Another example is the transfer of analytes from their original sample into a solvent that is more suitable for testing by GC. The injection of water into most GC columns creates problems with the long-term behavior and reproducibility of these columns. These problems occur because water tends to bind strongly to such columns, changing their chromatographic properties from one injection to the next. In addition, the water may contain dissolved solids, salts, or other nonvolatile compounds that are not suitable for injection into a GC system. Because of these problems, aqueous-based samples are usually extracted before GC analysis. Either liquid−liquid or solid-phase extraction can be used to remove analytes from water and place them into a more volatile solvent, as well as to remove them from nonvolatile substances present in the sample. The use of solvent evaporation with these two methods also makes it possible to greatly reduce the sample volume, which allows a larger mass of analyte to be injected for easier detection.

Headspace analysis is another technique used to avoid the introduction of water and nonvolatile compounds into a GC system. This technique is based on the fact that volatile chemicals in a liquid or solid sample will also be present in the vapor phase located above the sample. By sampling this vapor (known as the "headspace"), a portion of the volatile substances can be collected without interference from other less volatile compounds. This approach can be performed in one of two ways: the static and dynamic methods. In the static method, the sample is placed in a closed container, and its contents are allowed to distribute between the sample and its vapor phase. After equilibrium has been reached, a portion of the vapor phase is injected into the GC system. In the dynamic method, an inert gas is passed through the liquid sample, which carries away volatile compounds. This gas is then passed through a cold trap or solid adsorbent, which collects and concentrates the volatile solutes for later analysis.

GC is not limited to gas and liquid samples but can be adapted for work with chemicals adsorbed or held within a solid sample. A common way of handling this type of analysis is to first extract the compounds of interest from the solid material by using liquid−liquid extraction, with the extracted analytes and their solvent then being treated as a liquid sample. It is also possible to analyze some solids without performing an extraction. For instance, headspace analysis might be used to examine the volatile compounds that are present in a solid. Another approach for examining solids is to use pyrolysis−GC. This approach is useful in forensic testing for solid substances like plastics and polymers that are not volatile and cannot be easily derivatized into a volatile form. Pyrolysis involves heating a solid sample in a controlled fashion to very high temperatures in order to break the solid apart into smaller, more volatile fragments that can be examined by GC. The result is a gas chromatogram which provides a fingerprint of the volatile compounds that are given off as the test substance is heated.

Liquid chromatography

General terms and concepts

LC is a chromatographic technique in which the mobile phase is a liquid [2−4,7−9]. There are many ways LC can be performed, but most LC methods for clinical analysis use a system like the one shown in Fig. 8.3. This

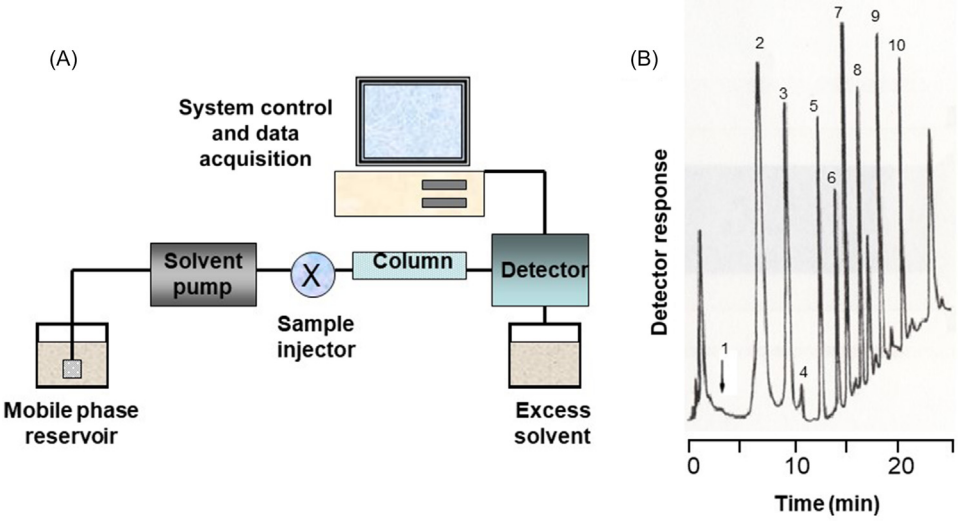


FIGURE 8.3 (A) A system for performing high-performance liquid chromatography and (B) a typical separation performed by high-performance liquid chromatography, using the analysis of basic drugs (peaks 1−10) in blood as an example. *The chromatogram in (B) is reproduced with permission from W.E. Lambert, D. Meyer, A.P. De Leenheer, Systematic toxicological analysis of basic drugs by gradient elution of an alumina-based HPLC packing material under alkaline conditions, J. Anal. Toxicol. 19 (1995) 73 [9]; courtesy Oxford University Press.*

system is known as a "liquid chromatograph," which includes a support and stationary phase enclosed in a column and a liquid mobile phase that is delivered to the column by means of a pump. An injection device is used to apply samples to the column, and a detector is usually present to detect and measure analytes as they leave the column.

Although LC and GC have many things in common, they also have some differences in terms of their sample and analyte requirements, their formats, and the role played by the mobile phase in these methods. The first requirement that must be met before a chemical can be examined by LC is that it must be possible to place this chemical into a liquid that can be injected into the mobile phase. This is a much less stringent requirement than the one presented in GC, where analytes must be naturally volatile or convertible to a volatile form. This feature makes LC valuable in separating many large biological compounds (e.g., proteins, peptides, and nucleic acids) that cannot be easily placed into the gas phase. The use of a liquid as the mobile phase also allows LC to be performed at much lower temperatures than GC, making it better suited for compounds that are thermally unstable.

A second requirement in LC is that there must be a difference in retention between the analytes to be separated. Although retention is altered in GC by adjusting the temperature and type of stationary phase within the column, retention in LC can also be varied by changing the mobile phase due to the higher density of liquids compared with gases. As a result, solute retention in LC depends on the interactions of sample components with both the mobile phase and the stationary phase, making LC more flexible than GC when optimizing a separation.

One disadvantage of LC is it tends to have greater band-broadening than GC. This greater band-broadening is due to the much slower diffusion coefficients of solutes in liquids versus gases. Thus, if it is possible to separate a series of chemicals by either LC or GC, GC is often the better choice, because it will provide narrower peaks and better separations. However, as discussed previously, the use of GC requires that the analytes be volatile or capable of being converted into a volatile derivative, which is not needed in LC. The greater density and viscosity of liquids compared with gases means that higher pressures are needed in LC than in GC to deliver mobile phase to the column. As a result, columns in LC are often much shorter than those used in GC, which also tends to make the number of theoretical plates lower for LC columns than for GC columns.

One result of the greater band-broadening and shorter columns in LC is that more emphasis is placed on obtaining more efficient supports with less band-broadening. Prior to the mid-1960s all LC columns contained supports made of relatively large and irregularly shaped particles. The use of such supports created broad peaks and separations with low resolution. This approach is sometimes called "classical LC" and is still commonly used for sample pretreatment because of its low cost and ease of use. However, a shift then occurred in LC toward the use of smaller and more uniform and rigid supports. This shift led to a technique known as high-performance liquid chromatography (HPLC). The presence of a more efficient support in this method produces narrower peaks, which in turn gives rise to better separations and lower limits of detection. These features make HPLC the current LC method of choice for clinical analysis. One consequence of using smaller support particles in HPLC is the need for greater pressures to pass the mobile phase through the column. Most modern HPLC columns require operating pressures of a few hundred to a few thousand pounds per square inch. This calls for special pumps and other system components that can be operated at such pressures (see Fig. 8.3). The sample in HPLC analysis is applied by using a closed system (e.g., an injection valve), and detection is typically performed by using an on-line detector.

Liquid chromatography mobile phases and elution methods

An important difference between LC and GC is that the retention of solutes in LC depends on interactions involving both the mobile and stationary phases. To describe how solutes are retained on an LC column in the presence of a given liquid, the terms "weak mobile phase" and "strong mobile phase" are used. A strong mobile phase is a solvent or a solution that quickly elutes a solute from a column. This situation occurs when the mobile phase is very similar to the stationary phase in the types of interactions that each can have with solutes. A weak mobile phase is a solvent or solution that elutes solutes slowly from a column, which occurs if the mobile phase is very different from the stationary phase in its interactions with solutes. It is important to note that whether a liquid is a weak or strong mobile phase will depend on the type of stationary phase in the column. For instance, a nonpolar solvent (such as hexane) will be a weak mobile phase on a polar stationary phase but a strong mobile phase for a nonpolar stationary phase. Thus it is the stationary phase that determines what liquids are strong or weak mobile phases for a given LC column.

Analytes can be eluted from an LC column by using either constant conditions or by using gradient elution. If a constant mobile phase composition is used for elution, this technique is referred to as "isocratic elution." Although this type of elution is simple and inexpensive, it does make it difficult to elute all solutes with good

resolution and in a reasonable amount of time. An alternative approach is to change the composition of the mobile phase with time. This approach is known as "solvent programming." This method is conducted by starting with a weak mobile phase, which allows early eluting solutes to be more strongly retained by the column. A switch is then made to a stronger mobile phase to allow highly retained solutes to come out in a reasonable amount of time. Solvent programming can be performed in one or more steps and by using a linear or nonlinear change in the mobile phase content, with linear gradients between two solvents being the most common method.

Temperature programming can also be used in LC, but it is much less common here than in GC. Temperature programming does improve resolution in some types of LC, such as size-exclusion chromatography (SEC) and ion-exchange chromatography (IEC). However, this method gives only marginal improvements in other types of LC and can cause degradation for some solutes. Flow programming can be used in LC as well. Advantages of flow programming include its ability to more quickly get the LC system ready for the next injection than solvent or temperature programming. A disadvantage of using flow programming in LC is the increased pressure across the column that occurs with an increase in flow-rate.

Types of liquid chromatography

Adsorption chromatography

A common way of grouping LC techniques is according to the mechanisms by which they separate solutes. The first of these is adsorption chromatography, which is a chromatographic technique that separates solutes based on their adsorption to the surface of a support. This method is also known as LSC. This technique is similar to GSC in that it uses the same material as both the stationary phase and the support. In fact, many of the supports used in GSC are also used in LSC.

The solid supports used in adsorption chromatography may be either polar or nonpolar in nature. Silica is the most popular support for adsorption chromatography and is a polar material with slightly acidic properties. Because silica is polar in nature, it will have the highest retention for polar compounds and will have a strong mobile phase that is polar in nature. Alumina is another popular support employed in adsorption chromatography. It is also polar and has the highest retention for polar compounds. Charcoal is also sometimes used as a support in adsorption chromatography. Unlike silica or alumina, charcoal is nonpolar and most strongly retains nonpolar solutes. The strong mobile phase on a charcoal column will also be nonpolar.

Adsorption chromatography is particularly useful in separating compounds with polar groups but that are soluble in nonpolar solvents; examples include steroids such as estrogens and testosterone. This method is also valuable in the separation of geometrical isomers and structurally similar chemicals. There are, however, problems in the use of adsorption chromatography. These problems include the heterogeneous nature of the surface on silica or alumina and the ability of these surfaces to act as catalysts for some chemical reactions. These supports can also give rise to nonreproducible retention for polar compounds and require the use of good quality solvents to give consistent mobile phase strengths.

Partition chromatography

The second major type of LC is partition chromatography. This is a chromatographic technique in which solutes are separated based on their partitioning between a liquid mobile phase and a stationary phase coated on a solid support. The support used in partition chromatography is usually silica but can also be other materials. Originally, partition chromatography involved coating the support with a liquid stationary phase that was immiscible with the mobile phase. However, most modern columns for partition chromatography employ stationary phases that are chemically bonded to the support.

There are two main types of partition chromatography. The first of these is normal-phase chromatography [also called "normal-phase liquid chromatography" (NPLC)], which is a type of LC in which the stationary phase is polar. These stationary phases contain polar regions based on cyano, amino, or diol groups that can form hydrogen bonds or undergo dipole interactions with solutes. Because NPLC has a polar stationary phase, it has the strongest retention for polar compounds. However, it may also be used for separating nonpolar compounds. The weak mobile phase in NPLC is a nonpolar liquid, such as an organic solvent, which is used as the injection solvent. A strong mobile phase is a polar liquid, such as water or methanol.

Reversed-phase chromatography [also known as "reversed-phase liquid chromatography" (RPLC)] is the second type of partition chromatography. It uses a nonpolar stationary phase, which is opposite or "reversed" in polarity from that utilized in normal-phase chromatography. This stationary phase usually consists of an alkane like an octyl group (C_8) or octadecyl group (C_{18}). RPLC is currently the most popular type of LC. The main reason for its popularity is that its weak mobile phase is a polar solvent, such as water. This feature makes RPLC ideal for the injection and separation of solutes in aqueous-based systems, such as drugs in clinical samples.

NPLC has similar applications to adsorption chromatography performed with silica or alumina supports. These applications typically involve the use of NPLC for separating chemicals in organic solvents and that contain polar functional groups. Examples include steroids, pesticides, terpenoids, nonionic detergents, sugars, and metal complexes. RPLC is used to separate a broad range of substances in aqueous samples, including its use for both small molecules (such as drugs) and macromolecules (such as proteins).

Another type of partition chromatography is hydrophilic interaction liquid chromatography (HILIC). Like NPLC, HILIC uses a polar stationary phase. However, in HILIC, the components of a sample now distribute between an organic-rich region in the mobile phase and a more polar water-enriched layer located at or near the surface of the support. The polar groups on the support's surface, which can be neutral or charged, may also interact with chemicals as they enter the water-enriched layer on the support. HILIC has been of particular interest for the analysis of polar compounds in fields such as proteomics and glycomics.

Ion-exchange chromatography

Another major type of LC is IEC. This is a chromatographic technique in which solutes are separated by their adsorption onto a support containing fixed charges on its surface. IEC is a fairly common technique used for the removal or replacement of ionic compounds in samples. It is also used for the separation and analysis of charged compounds, including inorganic ions, organic ions, and biological compounds, such as amino acids, proteins, and nucleic acids.

There are two types of stationary phases used in IEC. The first type is a cation-exchanger, which has a negatively charged group and is used to separate positively charged ions. The second type is an anion-exchanger, which has a positively charged group and is used to separate negatively charged ions. These two types of stationary phases are used in the methods of cation-exchange chromatography and anion-exchange chromatography, respectively. There are several types of stationary phases used in these methods. For cation-exchange chromatography, the charged groups consist of either: (1) the conjugate base of a strong acid (e.g., a sulfonic acid group), which is ionized over a broad pH range; or (2) the conjugate base of a weak acid (e.g., a carboxylate group), which has a net charge over a moderate pH range that is above the pK_a of the weak acid. For anion-exchange chromatography, the stationary phase is either: (1) the conjugate acid of a strong base (e.g., a quaternary amine), which is ionized over a broad pH range; or (2) the conjugate acid of a weak base (e.g., a tertiary amine), which has a net charge over a moderate pH range that is below the pK_a of its conjugate acid.

Several supports can be used with these cation- or anion-exchange groups. One common support used in IEC for small inorganic and organic ions is polystyrene. Carbohydrate-based gels are another common type of support used in IEC. Examples of these include agarose, cross-linked dextran, and cellulose gels. Silica can also be used as a support in IEC. Each of these supports is modified prior to placing appropriate charged groups on their surface. The high efficiencies and rigid structure of silica make it useful for HPLC-based IEC. The main limitation of silica in these applications is the narrower range of pH values over which it is stable compared with polystyrene- or carbohydrate-based supports.

A strong mobile phase in IEC is a solution with a high concentration of competing ions. A change in ion concentration is the most common way of altering the retention of sample ions in IEC. However, the retention of charged analytes will also be affected by the type of competing ion, the type of ion-exchange site being used and the mobile phase pH, the latter of which can alter the charge on some analytes, competing ions, or ion-exchange sites. Adding a complexing agent to the mobile phase can also affect the charge of an analyte and alter its retention. For instance, complexation of the cation Fe^{+3} with excess Cl^- can be used to form a negatively charged $FeCl_4^-$ complex, which can then be retained and analyzed by anion-exchange chromatography.

As a preparative tool, IEC has been used for many years in biochemistry for purifying proteins, peptides, and nucleotides. Ion-exchange supports are also frequently employed for concentrating small inorganic and organic ions from food and environmental samples. In addition, IEC has been used for many years for direct chemical analysis. One common example in clinical chemistry is the use of IEC to separate amino acids and amines to look for metabolic disorders in patients, as illustrated in Fig. 8.4.

The use of a high concentration of competing ions makes it difficult to detect cations or anions as they leave ion-exchange columns. A way to overcome this problem is to use a special type of IEC known as "ion chromatography" (IC). In this method, the background signal due to competing ions is reduced by using a low number of charged sites for the stationary phase. The result is that a lower concentration of competing ions will be needed to elute sample ions. This method is often used with a second column or membrane separator (of opposite charge to the first ion-exchange column), in which competing ions that have high conductivity are replaced with chemicals or ions that have lower conductivity. The result is a lower signal due to the competing ions, providing better limits of detection for analyte ions.

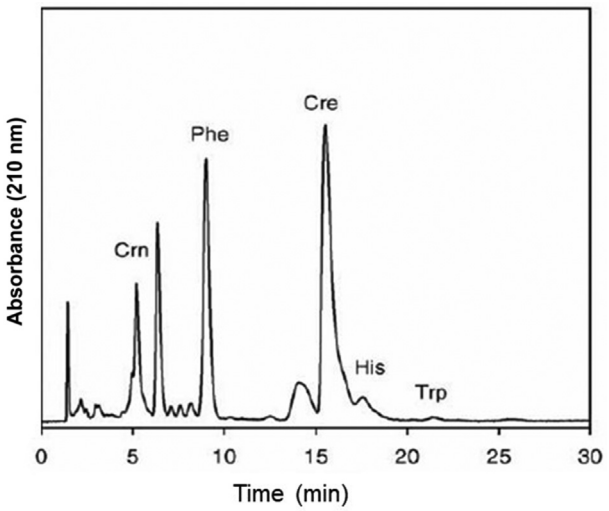


FIGURE 8.4 Use of ion-exchange chromatography to separate and analyze UV-absorbing amino acids in urine for an individual with phenylketonuria. Abbreviations: *Cre*, creatinine; *Crn*, creatine; *His*, histidine; *Phe*, phenylalanine; *Trp*, tryptophan. *Reproduced with permission from Y. Yokoyama, K. Yamasaka, H. Sato, Simultaneous determination of urinary creatinine and UV-absorbing amino acids using a novel low-capacity cation-exchange chromatography for the screening of inborn errors of metabolism, J. Chromatogr. B 816 (2005) 333 [10]; courtesy Elsevier.*

Size-exclusion chromatography

The fourth general type of LC is size-exclusion chromatography (SEC). This is a liquid chromatographic technique that separates substances according to their size. This technique is based on the different ability of analytes to access mobile phase within the pores of a support. No true stationary phase is present in this system. Instead, the mobile phase in the pores acts as the stationary phase. SEC uses a support that has a certain range of pore sizes. As solutes travel through this support, small molecules can enter the pores while large molecules cannot. Larger molecules can enter a smaller volume of the column, so they elute before smaller molecules. The result is a separation based on size or mass.

The ideal support in SEC consists of a porous material that does not interact directly with the injected solute. Carbohydrate-based supports like dextran and agarose can be used in SEC for biological compounds and aqueous-based samples. Polyacrylamide gel can also be employed for such samples. Similarly, polystyrene can be used for SEC when working with samples in organic solvents, and silica containing a diol-bonded phase can be utilized with aqueous samples.

The mobile phase in SEC can be either a polar or nonpolar solvent. Because there is no true stationary phase, there is also no weak or strong mobile phase in this method. Instead, the selection of mobile phase depends mostly on the solubility of the analytes and the support's stability. If an aqueous mobile phase is used in SEC, the technique is called "gel filtration chromatography." If an organic mobile phase is used, the technique is known as "gel permeation chromatography." As a preparative tool, SEC is often used with biological samples to remove small solutes from large agents like proteins. It can also be used to transfer large analytes from one solution to another or to remove salts from a sample. This approach can further be used to estimate the molecular weight of an analyte like a protein by comparing the retention of such an analyte to that of similar standard compounds with known masses.

Affinity chromatography

The fifth type of LC is affinity chromatography (AC). This method is based on biologically related interactions, like the binding of an enzyme with its substrate, or the binding of an antibody with an antigen. These reactions are used in AC by immobilizing one of a pair of interacting molecules onto a solid support and placing it into a column. This immobilized molecule is known as the "affinity ligand," and it is this agent that acts as the stationary phase in this method. The column containing the immobilized ligand can then be employed as a selective adsorbent for the complementary molecule.

The sample in AC is often applied to the column in the presence of an application buffer during an adsorption step. Because of the strong and selective nature of most biological interactions, the affinity ligand will bind to the analyte of interest during this step while allowing most other sample components to pass through as a nonretained peak. After these nonretained components have been washed from the column, an elution buffer is applied to release the retained analyte. This analyte is then detected as it leaves the column or collected for later use. The column and affinity ligand are then placed back into the application buffer, allowing them to be regenerated prior to injection of the next sample.

There are several types of ligands employed in AC, but all can be classified into one of two categories: (1) high-specificity ligands and (2) general (group-specific) ligands. High-specificity ligands are compounds that bind to only one or a few very closely related molecules. Examples include antibodies for binding to foreign agents (i.e., antigens) and single-stranded nucleic acids for separating and binding to complementary strands. General, or group-specific, ligands are molecules that bind to a family or class of related molecules. Examples of these include boronates for binding compounds with diol groups, and lectins for retaining compounds that have carbohydrate groups.

Several types of supports are utilized in AC. Carbohydrate gels like agarose or cellulose are commonly

used with affinity ligands for the purification of biological molecules. Silica can also be used by first converting this support into a diol-bonded form. The affinity ligand is then attached to these supports by using an immobilization method. This process usually involves coupling the affinity ligand to the support through amine, carboxyl, or sulfhydryl groups. If appropriate immobilization conditions are not used, the ligand may be denatured or attached in a way that blocks its binding to the analyte. For the immobilization of small molecules, placement of a spacer arm between the ligand and the support may be necessary to reduce steric hindrance in the binding of larger molecules to this ligand.

A weak mobile phase in AC is one that allows strong binding between the analyte and the ligand. This weak mobile phase is usually a solvent that mimics the pH, ionic strength, and polarity of the affinity ligand in its natural environment. This mobile phase is known as the "application buffer." It is this solvent that is typically used during the application, washing, and regeneration steps. A strong mobile phase in AC is a solvent that can readily remove the analyte from the ligand. This is called the "elution buffer." This strong mobile phase is selected to either lower the association equilibrium constant for the analyte−ligand interaction or to displace the analyte from the affinity ligand through the use of a competing agent.

AC is frequently used as a large-scale purification method for enzymes and proteins. This application involves the use of columns that contain immobilized agents that can selectivity bind to and retain such substances in the presence of other sample components. However, AC is also commonly used as a method for sample preparation. Examples include the use of affinity columns containing antibodies for the isolation of cellular proteins or the use of immobilized metal ions to isolate recombinant proteins containing histidine tags as part of their structure. The selectivity of AC has also made it appealing for use in the direct analysis of complex biological samples. One example is the use of boronate affinity columns in the measurement of glycated hemoglobin, an indicator of long-term blood sugar levels in diabetes. Affinity columns have also been used with HPLC for the measurement of hormones, proteins, drugs, and other agents in clinical and biological samples.

Liquid chromatography detectors

As shown in Table 8.3, there are several types of detectors available for LC. These can again be classified as either general or selective detectors. An absorbance detector is an example of a general detector for LC. This type of device is similar to those used for other absorbance measurements but includes a special sample cell (known as a "flow cell"), which allows the mobile phase and analytes to pass through in a continuous manner. The simplest type of absorbance detector for LC is a fixed-wavelength absorbance detector. This device is set to always monitor a specific wavelength (usually 254 nm). A variable-wavelength absorbance detector has a more flexible design and allows the monitored wavelength to be varied over a wide range (e.g., 190−900 nm). A third design is a photodiode array detector, which uses an array of small detector cells to measure the change in absorbance at many wavelengths simultaneously. This array makes it possible to record an entire spectrum for an eluting compound. Absorbance detectors can detect any compound that absorbs light at the wavelength(s) monitored. They can also be used with gradient elution, provided that the weak and strong mobile phases do not have significant

TABLE 8.3 Common detectors for liquid chromatography.

Detector name	Compounds detected	Detection limit[a]
Refractive index detector	General: all compounds	0.1−1 μg
UV−vis absorbance detector	Compounds with chromophores	0.1−1 ng
Evaporative light scattering detector	Nonvolatile compounds	0.1 μg
Fluorescence detector	Fluorescent compounds	1−10 pg
Conductivity detector	Ionic compounds	0.5−1 ng
Electrochemical detector	Electrochemically active compounds	0.01−1 ng
Mass spectrometry	General: full-scan mode	0.1−1 ng
	Selective: SIM mode	

SIM, Selected ion monitoring.
[a]*This information is from Poole and Poole [2] and manufacturers of these detectors.*

differences in their absorbances at the detection wavelengths. The main disadvantage of these devices is they require a compound to have a chromophore that can absorb at the wavelengths being monitored or that can be derivatized into a form that does absorb.

A refractive index (RI) detector is one of the most universal detectors available for LC. This detector measures the ability of the mobile phase and analytes to refract or bend light. This property varies as the composition of the mobile phase changes, such as when analytes elute from a column. An RI detector will respond to any compound that has a different RI from the mobile phase, provided that enough solute is present to give a measurable signal. This property makes an RI detector useful in work where an analyte cannot be easily measured by other devices or where the nature or properties of an analyte are not yet known. However, an RI detector does not have limits of detection as low as absorbance detectors or other LC detectors. In addition, its signal is sensitive to changes in the mobile phase composition and temperature, making it difficult to use with gradient elution.

A third type of general detector for LC is an evaporative light scattering detector (ELSD). This device can be used for any solute that is less volatile than the mobile phase and detects analytes by first converting the eluting mobile phase into a spray of small droplets. As the solvent in these droplets evaporates, small particles of the nonvolatile sample components are left behind that can scatter light. The extent of this scattering will depend on the number and size of these particles and the original concentration of nonvolatile solutes in the mobile phase. An ELSD has a better limit of detection than an RI detector and can be used with gradient elution. Although absorbance detectors have better limits of detection, an ELSD does not require that a chromophore be present in the analyte. This feature gives an ELSD the ability to examine substances that cannot be easily detected based on their absorbance, such as lipids and carbohydrates.

An example of a selective detector for LC is a fluorescence detector. This device measures the ability of solutes to absorb and emit light at a particular set of wavelengths. Fluorescence can also be used to detect compounds that are first converted to a fluorescent derivative, as is often used for analytes such as alcohols, amines, amino acids, and proteins. Although fluorescence detectors can be used with gradient elution, extremely pure mobile phases must be employed, because even trace impurities can affect the background signal or quench the fluorescence of solutes.

A conductivity detector can monitor ionic compounds in LC. This detector measures the ability of the mobile phase and its contents to conduct a current when placed in an electrical field. Conductivity detectors can be used to detect any compound that is ionic, making them useful in IC. This device can be used with gradient elution as long as the ionic strength (and possibly pH) of the mobile phase is kept constant. It is also necessary that the background conductance of the mobile phase be sufficiently low so that sample ions can be detected.

An electrochemical detector is another device used to monitor specific compounds in LC. This combination is known as "LC/electrochemical detection" (LC/EC). An electrochemical detector can be used to measure the ability of an analyte to undergo either oxidation or reduction. Examples of compounds that may be detected by reduction in such a detector include aldehydes, ketones, esters, unsaturated compounds, and aromatics. Compounds that may be detected by their oxidation include phenols, mercaptans, aromatic amines, diamines, purines, and dihydroxy compounds, such as some carbohydrates. The response of an electrochemical detector depends on the extent of oxidation or reduction that occurs at the given potential of the electrode. For such compounds, the limit of detection can be quite low due to the accuracy with which electrical measurements (especially current) can be made.

Another type of detector that can be employed in LC is a mass spectrometer. The result is a technique known as "liquid chromatography/mass spectrometry" (LC/MS). The use of mass spectrometry with LC can be used to both measure the amount of a chemical and identify this chemical based on its molecular ion and/or fragment ions. By looking at all ions that are produced in the mass spectrometer (i.e., the full-scan mode), LC/MS can be used as a general detection method. If LC/MS is instead used to look at a few ion characteristics of a particular set of analytes (i.e., the SIM mode), this approach can also be used for selective detection.

The most common way of performing LC/MS is to use electrospray ionization (ESI). This combination can examine substances ranging from small polar compounds to proteins. It can also be utilized with gradient elution methods. The use of LC/MS with ESI is particularly useful in the analysis of proteins and peptides, which tend to give ions with mass-to-charge ratios that are outside the range of most common mass analyzers. In ESI, this problem is overcome by the fact that many charges are often placed on one biomolecule, giving ions with lower mass-to-charge ratios. One difficulty associated with this process is that a single protein or peptide can give rise to many molecular ions. However, this issue can be addressed by using computer programs that are designed to analyze such spectra and determine the true molecular masses from such information.

Introduction to electrophoresis

General terms and concepts

Another separation method often used in clinical testing is electrophoresis [11]. This is a technique in which

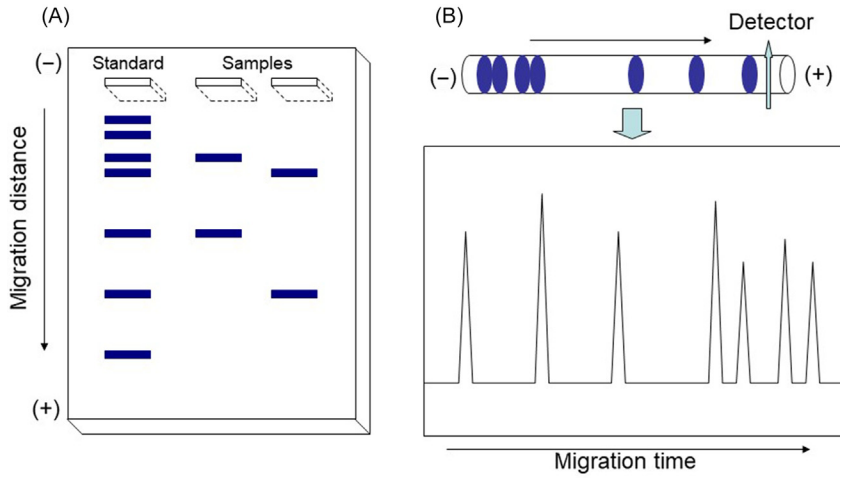


FIGURE 8.5 General scheme for the separation of chemicals by (A) gel electrophoresis and (B) capillary electrophoresis.

solutes are separated by their different rates of migration in an electric field. This concept is illustrated in Fig. 8.5, in which a sample is first placed in a container or support that also contains a background electrolyte (or "running buffer"). When an electric field is applied to this system, ions in the running buffer will flow from one electrode to the other and provide the current needed to maintain the applied voltage. At the same time, positively charged ions in the sample will move toward the negative electrode (the cathode), while negatively charged ions will move toward the positive electrode (the anode). The result is a separation of these ions based on their charge and size. Many biological compounds have charges or ionizable groups (e.g., DNA and proteins), which makes electrophoresis a common tool in clinical laboratories and medical research.

Electrophoresis can be performed in a variety of formats. One way is to apply small amounts of a sample to a support (usually a gel) and allow analytes to travel in a running buffer through the support when an electric field is applied. This approach is known as "gel electrophoresis." It is also possible to separate the components of a sample in a narrow capillary filled with a running buffer and placed into an electric field. This second method is called "capillary electrophoresis" (CE).

Depending on the type of electrophoresis being performed, the result of the separation can be viewed in one of two ways. In the case of gel electrophoresis, the separation is stopped before the analytes have traveled off the support. The result is a series of bands in which the migration distance characterizes the extent these substances have interacted with the electric field. Because the migration distance of an analyte through an electrophoresis gel will depend on the exact voltage and time used for the separation, it is common to include standard samples on the support to help in analyte identification. The intensities of the resulting bands are then used to measure the amounts of analytes in the sample.

In CE, all analytes travel the same distance, from the point of injection to the opposite end, where a detector is located. However, the analytes now differ in the time it takes to travel this distance. In this case, it is the migration time for each analyte that is measured and recorded. The resulting plot of detector response versus migration time is called an "electropherogram." This migration time is used for analyte identification, while the peak height or peak area is utilized for measurement. An internal standard is usually injected along with the sample to correct for variations during injection or for small fluctuations in the experimental conditions during the separation.

Theory of electrophoresis

The separation of analytes by electrophoresis requires two items. First, there must be a difference in how the analytes will interact with the electric field. This requirement means the analytes must have a difference in their migration times or migration distances. The second requirement is that the bands or peaks for the analytes must be sufficiently narrow to allow them to be resolved.

The overall rate of travel of a charged solute in electrophoresis will depend on two opposing forces. The first of these is the attraction of a charged solute toward the electrode of opposite charge. This force depends on the strength of the applied electric field (E) and the charge on the solute (z). However, the solute will also experience resistance to its movement by the surrounding medium. The force of this resistance depends on the size of the solute (as described by the solvated radius r), the viscosity of the medium (η), and the velocity of the solute (v). When an electric field is applied, a solute will accelerate toward the electrode of opposite charge until these two

opposing forces become equal in size but opposite in direction. At this point, a steady-state situation is produced in which the solute moves at a constant velocity, as given by Eq. (8.9).

$$v = \frac{(Ez)}{(6\pi r\eta)} \tag{8.9}$$

To see how this velocity will be affected by only the strength of the electric field, other terms in Eq. (8.9) can be combined to give a single constant (μ),

$$v = \mu E \tag{8.10}$$

where $\mu = z/(6\pi r\eta)$. This combination of terms is known as the "electrophoretic mobility." The value of μ is constant for a given analyte under a particular set of temperature and solvent conditions. However, this value will depend on the apparent size and charge of the solute, as represented by the ratio z/r in Eq. (8.9). As a result, any two solutes with different charge-to-size ratios can, in theory, be separated by electrophoresis.

It is also possible to have movement of the running buffer in an electric field. This movement can occur if there are any fixed charges present within the system, such as on the surface of a support. The presence of these fixed charges attracts ions of opposite charge from the running buffer and creates an electrical double layer at the support's surface. In the presence of an electric field, this double layer acts like a piston that causes a net movement of the buffer toward the electrode of opposite charge to the fixed ionic groups. This process is known as "electroosmosis" and results in a net flow of the buffer within the system.

The extent to which electroosmosis affects the buffer and analytes in electrophoresis is described by the electroosmotic mobility (μ_{eo}). This term depends on such things as the size of the electric field, the type of running buffer, and the charge on the support. Depending on the direction of buffer flow, electroosmosis can work either with or against the natural migration of an analyte through the electrophoretic system. The overall observed electrophoretic mobility (μ_{Net}) for an analyte will equal the sum of its own electrophoretic mobility (μ) and the mobility of the running buffer due to electroosmotic flow (μ_{eo}).

$$\mu_{Net} = \mu + \mu_{eo} \tag{8.11}$$

In gel electrophoresis, electroosmotic flow is usually small compared with the inherent rate of analyte migration. However, this is not usually the case in CE, where the support has a relatively large charge and high surface area compared with the volume of running buffer.

The same terms used to describe band-broadening in chromatography (e.g., the number of theoretical plates N and the height equivalent of a theoretical plate H) can be used in electrophoresis. However, chromatography and electrophoresis differ in which processes are most important in producing this band-broadening. For instance, one important band-broadening process in many types of electrophoresis is *Joule heating*. This is heating that occurs whenever an electric field is applied to the system. According to *Ohm's law*, placing a voltage V across a medium with a resistance of R requires that a current of I be present to maintain this voltage.

$$V = IR \tag{8.12}$$

However, as current flows through the system, heat is generated. This heat production depends on the voltage, current, and time t that current passes through the system, as shown below.

$$\text{Heat} = VIt \tag{8.13}$$

As heat is produced, the temperature of the electrophoretic system will begin to rise and increase the rate of diffusion. In addition, if the heat is not distributed uniformly, the temperature will not be the same throughout the system. This effect leads to regions with different densities, which causes mixing and results in even more band-broadening.

One way Joule heating can be decreased is by using a lower voltage for the separation. However, using a lower voltage will also lower the migration velocities of analytes and give longer separation times. An alternative approach is to use more efficient cooling, which allows higher voltages to be employed and provides shorter separation times. A third possibility is to add a support to minimize the effects of uneven heat distribution and density gradients in the running buffer. Another factor that affects Joule heating is the ionic strength of the running buffer, where a lower ionic strength buffer will decrease heat production.

Wick flow is a source of band-broadening that occurs in gel electrophoresis. In such a system, the gel is kept in contact with the electrodes and buffer reservoirs through the use of wicks. The gel is often open to air, which means Joule heating will cause evaporation of solvent in the running buffer at the surface of this support. As this solvent is lost, it is replenished by the flow of more solvent through the wick and from the buffer reservoir. This process leads to a net movement of buffer toward the center of the support. Because the rate of this flow depends on the rate of evaporation, it increases with high voltages and high currents.

Gel electrophoresis

General terms and concepts

One of the most common types of electrophoresis is the method of gel electrophoresis. This is an electrophoretic method that is performed by applying a sample to a gel

support that is then placed into an electric field. In this type of system, several samples are usually applied to the gel and allowed to migrate in the presence of an electric field. The separation is stopped before analytes leave the gel, with the location and intensities of their bands then being determined.

Supports and sample application in gel electrophoresis

A system for performing gel electrophoresis may involve a support held in either a vertical or horizontal position. The type of support used will depend on the analytes and samples being examined. Cellulose acetate and filter paper are valuable for work with small molecules like amino acids and nucleotides. Starch is also sometimes used as a support in electrophoresis. However, a more common support is agarose. In addition to its low nonspecific binding for many biological compounds and its low inherent charge, agarose gels have relatively large pores. This property makes agarose useful in dealing with large molecules like DNA, as occurs in DNA sequencing.

Probably, the most common support used in gel electrophoresis is polyacrylamide. This combination is often referred to as "polyacrylamide gel electrophoresis" (PAGE). Polyacrylamide is a synthetic polymer that can be made with a variety of pore sizes. These pores are generally smaller than those in agarose and of a size more suitable for the separation of proteins and peptide mixtures. Like agarose, polyacrylamide has low nonspecific binding for many biological compounds and does not have any inherent charged groups in its structure.

The samples in gel electrophoresis are applied by a micropipette to small wells that are made in the gel during its preparation. This process usually involves applying sample volumes in the range of 10−100 μL. These sample volumes help provide a sufficient amount of analytes for later detection and collection. However, there is also a danger of introducing band-broadening by creating a large sample band at the beginning of the separation. A common approach to create narrow sample bands is to employ two types of gels in the system: a stacking gel and a running gel. The stacking gel has a low degree of cross-linking (giving it large pores) and is located in the section of the support in which the sample wells are located. The running gel has a higher degree of cross-linking (i.e., smaller pores) and is used during the actual separation of analytes. After a sample has been placed in the wells in the stacking gel, analytes will travel quickly through this medium until they reach its boundary with the running gel. These substances will then travel much more slowly, allowing other parts of the sample to catch up and form a narrower, more concentrated band that can then be separated into its components as they travel through the running gel.

Detection in gel electrophoresis

There are several ways in which analytes can be detected in gel electrophoresis. This detection is performed after the analytes have been separated either by looking at their location on the gel or by transferring these substances to a different support for detection. In some cases, it is possible to perform direct detection of analytes on a gel. Direct detection can sometimes be conducted visually when dealing with intensely colored proteins, like hemoglobin, or by using absorbance measurements and a scanning device called a densitometer. However, direct detection requires that an analyte be present at a relatively high level in a sample or concentrated by a pretreatment step prior to the separation.

The most common approach for detection in gel electrophoresis is to treat the support with a stain or reagent that makes it easier to see analyte bands. Examples of stains used for proteins are Amido black, Coomassie Brilliant Blue, and Ponceau S. Silver nitrate, giving rise to a method known as "silver staining," is also used when looking at low concentration proteins. DNA bands are often detected by using ethidium bromide. When separating enzymes, the natural catalytic ability of these substances can be used for their detection (e.g., detecting production of the fluorescent compound NAD(P)H by enzymes that generate this substance in their reactions). Sometimes biological ligands are used to react with analytes on a gel for detection. This approach is used in rocket immunoelectrophoresis, where antibodies are used to form precipitation bands that help identify and measure specific proteins in a sample.

A third approach for detection in gel electrophoresis is to transfer a portion of the analyte bands to a second support (e.g., nitrocelluose), where they are reacted with a labeled agent. This approach is known as "blotting." There are several different types of blotting methods. A Southern blot is used to detect specific sequences of DNA by having these sequences bind to an added, known sequence of DNA containing a radioactive tag (^{32}P) or other easily detected label. A Northern blot is similar in format but is instead used to detect specific sequences of RNA by using a labeled DNA probe. A Western blot is used for detecting a given protein by reacting this protein with labeled antibodies that can bind to such analytes.

Special types of gel electrophoresis

Whenever there is a porous support in an electrophoretic system, it is possible that analytes may be separated by size as well as their electrophoretic mobilities. This effect

is used for proteins in a technique known as "sodium dodecyl sulfate polyacrylamide gel electrophoresis" (SDS-PAGE). In SDS-PAGE, the proteins in a sample are first denatured and their disulfide bonds are broken through the use of a reducing agent. This pretreatment converts the proteins into a set of single-stranded polypeptides. These polypeptides are then treated with sodium dodecyl sulfate (SDS). This is a surfactant with a nonpolar tail and a negatively charged sulfonic group at the other end. The nonpolar end of this surfactant coats the protein, forming a roughly linear rod with a layer of negative charges on the outside. The result for a mixture of proteins is a series of rods with different lengths but similar charge-to-mass ratios. Next, these protein rods are passed through a porous polyacryamide gel in the presence of an electric field. The negative charges on these rods from the SDS coating cause them all to move toward the positively charged electrode, while the pores of the gel cause small rods to travel more quickly than larger ones. At the end of the run, the positions of the protein bands from a sample are compared with those obtained in a protein ladder for a series of protein standards with known molecular masses and applied to the same gel, as illustrated in Fig. 8.5A. This comparison can be performed either qualitatively or by preparing a calibration curve by plotting the log of the molecular masses for the protein standards versus their migration distances.

Another type of electrophoresis that often makes use of supports is isoelectric focusing (IEF). This is a method used to separate zwitterions (i.e., substances with both acidic and basic groups) based on differences in their isoelectric points. This type of separation is accomplished by having the zwitterions migrate in an electric field across a pH gradient. In this pH gradient, each zwitterion will migrate until it reaches a region where the pH is equal to its isoelectric point (pI). Once this situation has occurred, the zwitterion will no longer have any net charge, and its electrophoretic mobility will become zero, causing the analyte to stop migrating. This results in a series of tight bands, where each band appears at the point where the pH is equal to the pI for a given analyte. IEF is useful in separating proteins with similar properties but slightly different amino acid compositions. A common example is its use in separating the isoforms of an enzyme. To obtain a separation in IEF, it is necessary to have a stable pH gradient. This pH gradient is produced by placing in the electric field a mixture of small reagent zwitterions known as "ampholytes." These are usually polyprotic amino carboxylic acids with a range of pK_a values. When these ampholytes are placed in an electric field, they will travel through the system and align in the order of their pK_a values. The result is the formation of a pH gradient. This gradient can be used directly or by first cross-linking the ampholytes to a support to keep them stationary in the system.

Another way gel electrophoresis can be utilized is in a method known as "two-dimensional electrophoresis." This is a high-resolution technique used to look at complex protein mixtures. In this method, two types of electrophoresis are performed on a single sample. The first of these separations is usually based on IEF, and the second is based on SDS-PAGE. The fact that two different characteristics of each protein are used in this separation (pI values and size) makes it possible to resolve a much larger number of proteins than is possible by either IEF or SDS alone. This method is popular in the analysis of large protein libraries, as is used for finding new disease biomarkers or in studying disease pathology.

Capillary electrophoresis

General terms and concepts

Another type of electrophoresis is the method of CE. This is a technique that separates analytes by electrophoresis that is carried out in a capillary. This method is typically conducted in capillaries with inner diameters of 20−100 μm and lengths of 20−100 cm. The use of these narrow bore capillaries allows efficient removal of Joule heating, which helps decrease band-broadening and provides much more efficient and faster separations than gel electrophoresis.

If there is no appreciable Joule heating present and no interactions between the analytes and the capillary, the main source of band-broadening in CE will be longitudinal diffusion. Under these conditions, an increase in applied voltage results in less band-broadening, because shorter times are allowed for analytes to undergo diffusion along the length of the capillary. Besides providing efficient separations, the ability to use high electric fields also reduces the time needed for a separation. This combination of good efficiency and speed has made CE of interest for the analysis of complex biological samples like urine and blood.

An example of a CE system is shown in Fig. 8.6. In addition to the capillary, this system includes a power supply and electrodes for applying the electric field, two buffer containers that supply a contact between these electrodes and the solution within the capillary, an on-line detector, and a means for injecting samples onto the capillary. Most CE instruments can supply voltages up to 25−30 kV. To prevent injury to laboratory workers, these instruments have safety features that isolate the region of high voltage from the user and that turn off this voltage when the system is opened for maintenance or for the insertion of samples and reagents.

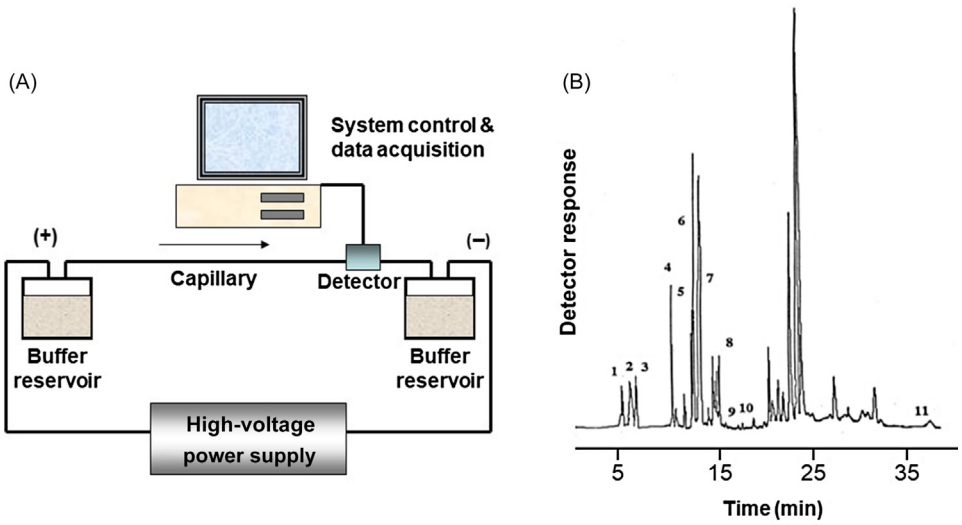


FIGURE 8.6 (A) A system for performing capillary electrophoresis and (B) a typical separation performed by capillary electrophoresis, using the analysis of biogenic amines (peaks 1–11) in urine. *The electropherogram in (B) is reproduced with permission from H. Siren, U. Karjalainen, Study of catecholamines in patient urine samples by capillary electrophoresis, J. Chromatogr. A 853 (1999) 527 [12]; courtesy Elsevier.*

Supports and sample application in capillary electrophoresis

The capillary in a CE system is usually fused silica. A fused silica capillary is often used directly, but it can also be modified to place various coatings on its interior surface. With a fused silica capillary, flow of the buffer due to electroosmosis can have a significant impact on the overall mobility of analytes. This electroosmosis is created by the negative charge present on surface of silica at a neutral or basic pH. The presence of these negatively charged groups creates an electrical double layer that causes movement of the running buffer toward the negatively charged electrode.

One useful feature of electroosmosis is it causes most analytes, regardless of their charge, to travel in the same direction through the CE capillary. This effect means that a sample containing many types of ions can often be injected at one end of the capillary (at the positively charged electrode), with electroosmosis then carrying these through to the other end (to the negatively charged electrode) and past a detector. This format is referred to as the "normal polarity mode" of CE. It is important to remember in this situation that a separation of ions will still occur but that the observed mobility will now be the sum of the inherent electrophoretic mobility for an analyte plus the mobility due to electroosmosis. The observed migration time for the analyte and the efficiency and resolution obtained for the separation are also affected.

Although many analytes will travel in the same direction as electroosmotic flow through a CE system, it is possible for some to have migration rates faster and in the opposite direction from electroosmosis. The analysis of these ions in a silica capillary is performed by injecting them by the negatively charged electrode and allowing them to migrate toward the positively charged electrode against electroosmotic flow. This format is referred to as the "reversed polarity mode" of CE. In addition, electroosmotic flow can be altered by changing the pH, which changes the charge on silica, or by placing a coating on the support surface. In this latter case, a neutral coating helps to reduce or eliminate electroosmosis while a positively charged coating will reverse the direction of this flow toward the positively charged electrode.

There are two features of CE that place special demands on how samples can be injected in this method. First, the small volume of this device must be considered. For instance, a typical 50-μm I.D. × 25-cm-long capillary for CE will contain only 0.5 μL of running buffer. This type of capillary requires the use of much smaller injection volumes to allow samples to be separated into their individual components. Another factor to consider is the high resolution of CE, which further restricts the sample sizes that can be employed. In practice, this factor generally requires the use of samples that are $<2\%$ of the capillary's internal volume, or <10 nL for a 0.5-μL volume capillary.

Sample volumes of nanoliters or less cannot be easily delivered by the types of injection valves used for LC. Thus alternative means for injection must be employed. One is hydrodynamic injection, which uses a difference in pressure to deliver a sample to the capillary. Hydrodynamic injection can be performed by placing one end of the capillary into a sample in an enclosed chamber and applying a pressure to this chamber for a fixed period of time. The amount of injected sample will depend on the size of the pressure difference and the amount of time this pressure is applied to the capillary. A similar effect occurs if a vacuum is pulled at the other end of the capillary. Once the sample has entered the capillary, the capillary is placed back into contact with the running buffer and electrodes, and the separation is begun.

Another technique frequently employed in CE is electrokinetic injection. This method again begins by placing the capillary into the sample; however, an electrode is also now placed into the sample. When an electric field is placed across the capillary, electroosmotic flow causes analytes to enter the capillary. The amount of each analyte that is injected will depend on the rate of electroosmotic flow, the analyte's inherent electrophoretic mobility, the applied electric field, and the time over which this field is applied. The capillary is then placed back into a container of the running buffer, and another electric field is applied to separate the sample components. One disadvantage to this approach is it favors the injection of analytes with electrophoretic mobilities that move them into the capillary over those that are applied only through the presence of electroosmotic flow.

Like traditional gel electrophoresis, there are various methods for concentrating samples and providing narrow analyte bands in CE. Sample stacking is one approach for accomplishing this goal and occurs when the ionic strength (and therefore the conductivity) of the sample is less than that of the running buffer. When an electric field is applied, analytes will migrate quickly through the sample matrix until they come to the boundary between the sample and the running buffer. Because the running buffer has a higher ionic strength, this situation causes the rate of migration to slow down and the analytes to concentrate as they enter the running buffer.

Detection in capillary electrophoresis

Many of the same detection techniques used for LC are also used in CE. However, one important difference between detection in LC and CE is the need in CE for methods that can work with very small sample sizes. This difference is a result of the small injection volumes that are required in CE to avoid adding excessive band-broadening. Selective detection methods that work particularly well for this purpose are electrochemical and fluorescence detection. However, UV−vis absorbance, conductance, and mass spectrometry detection are also employed.

Special types of capillary electrophoresis

One useful feature of gel electrophoresis is the ability of some supports in this method to separate analytes based on size, as occurs for proteins in SDS-PAGE. The same effect can be obtained in CE by including an agent in the CE system that sieves the analytes, or separates them based on size. There are several ways of doing this. The first way is to place a porous gel in the capillary, like the polyacrylamide gels employed in SDS-PAGE. This approach is called "capillary gel electrophoresis." However, these gels are not always stable in the high electric fields used in CE and must frequently be replaced. A second approach is to add a large polymer to the running buffer that is not immobilized but that can entangle with analytes and alter their rate of migration. This technique provides a system with better reproducibility and stability than those using gels, because the polymer is continuously renewed as the running buffer passes through the capillary.

The approach using polymers to entangle and separate analytes based on their size is often used when performing DNA sequencing by CE. The process of DNA sequencing by CE typically makes use of the Sanger method, in which the DNA template sequence to be examined is mixed with a smaller DNA primer that binds to part of this sequence. This mixture is then combined with: (1) nucleotides that contain the bases cytosine (C), guanine (G), adenine (A), or thymine (T); (2) the enzymes that are needed to add these nucleotides to the primer and extend it to match the template; and (3) small amounts of labeled nucleotides (also containing C, G, A, or T) that will stop this extension. The DNA strands that have been formed are then separated by CE based on their size, or strand length. The type of labeled nucleotide that ends each sequence is then determined and used to provide the sequence of the original template.

Ordinary CE works well for separating cations and anions, but it cannot be used to separate neutral substances, which migrate as a single peak that travels with the electroosmotic flow. However, CE can be extended to such compounds by placing in the running buffer a charged agent that can interact with these substances. One way of accomplishing this is to employ micelles as additives, giving a method known as "micellar electrokinetic chromatography." A micelle is a particle formed by the aggregation of a large number of surfactant molecules, such as SDS. As indicated earlier, SDS has a long nonpolar tail attached to a negatively charged sulfate group. When the concentration of a surfactant like SDS reaches a certain threshold level (i.e., the critical micelle concentration), some of the surfactant molecules come together to form micelles. If this situation occurs in a polar solvent like water, the nonpolar tails of the surfactant will be on the inside of the aggregate (giving a nonpolar interior), while the charged groups at the other end will be on the outside by the solvent.

When micelles based on SDS are placed into the running buffer of a CE system, they will be attracted toward the positively charged electrode. If a sample with several neutral compounds is now injected, some of these neutral substances may enter the micelles and interact with their nonpolar interior. Although these neutral compounds normally travel with electroosmotic flow through the capillary, while they are in the micelles they travel with the

micelles in the opposite direction. This effect gives a separation of neutral compounds based on the degree to which they enter the micelles. Micelles can also alter the migration times for charged substances through a combination of partitioning and charge interactions between the analytes and the micelles.

References

[1] H.Y. Aboul-Enein (Ed.), Separation Techniques in Clinical Chemistry, Marcel Dekker, New York, 2003.

[2] C.F. Poole, S.K. Poole, Chromatography Today, Elsevier, Amsterdam, 1991.

[3] L.E. Schoeff, R.H. Williams (Eds.), Principles of Laboratory Instruments, Mosby, Boston, 1993.

[4] J. Cazes (Ed.), Ewing's Analytical Instrumentation Handbook, third ed., CRC Press, Boca Raton, 2004.

[5] R.L. Grob, E.F. Barry (Eds.), Modern Practice of Gas Chromatography, fourth ed., John Wiley, Hoboken, 2004.

[6] M.C. Haven, G.A. Tetrault, J.A. Schenken (Eds.), Laboratory Instrumentation, fourth ed., Van Nostrand Reinhold, New York, 1995.

[7] W.J. Lough, I.W. Wainer (Eds.), High Performance Liquid Chromatography, Blackie Academic, New York, 1996.

[8] D. Corradini, T.M. Phillips (Eds.), Handbook of HPLC, second ed., CRC Press, Boca Raton, 2011.

[9] W.E. Lambert, D. Meyer, A.P. De Leenheer, Systematic toxicological analysis of basic drugs by gradient elution of an alumina-based HPLC packing material under alkaline conditions, J. Anal. Toxicol. 19 (1995) 73.

[10] Y. Yokoyama, K. Yamasaki, H. Sato, Simultaneous determination of urinary creatinine and UV-absorbing amino acids using a novel low-capacity cation-exchange chromatography for the screening of inborn errors of metabolism, J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci. 816 (2005) 333.

[11] J.P. Landers (Ed.), Handbook of Capillary and Microchip Electrophoresis and Associated Microtechniques, CRC Press, Boca Raton, 2008.

[12] H. Siren, U. Karjalainen, Study of catecholamines in patient urine samples by capillary electrophoresis, J. Chromatogr. A 853 (1999) 527.

Self-assessment questions

1. Which part of a chromatographic system interacts with the components of a sample and delays the rate at which they travel through the system?
   a. the mobile phase
   b. the stationary phase
   c. the support material
   d. the detector
2. Which of the following statements is false?
   a. The retention factor is related to the retention time of a compound.
   b. The van Deemter equation shows how the retention of a compound changes with linear velocity.
   c. The resolution between two peaks will increase with their degree of retention.
   d. The retention factor is affected by the amount of stationary phase in the column.
3. Which of the following items is NOT desired for a good separation in chromatography?
   a. a large value for the number of theoretical plates, N
   b. a large value for the peak resolution, R_s
   c. a large value for the plate height, H
   d. a large value for the separation factor, α
4. What is the most common type of column used in clinical applications of GC?
   a. a packed column with a bonded stationary phase
   b. a packed column with a liquid stationary phase
   c. an open-tubular column with a bonded stationary phase
   d. an open-tubular column with a liquid stationary phase
5. A GC detector that is selective for compounds with electronegative groups is _____.
   a. a thermal conductivity detector
   b. a flame ionization detector
   c. a nitrogen−phosphorus detector
   d. an electron capture detector
6. Liquid chromatography is different from gas chromatography in that _____.
   a. it has less band-broadening
   b. it has a liquid mobile phase that affects analyte retention
   c. it requires volatile analytes
   d. it can use a gas or liquid as the mobile phase
7. Which type of liquid chromatography uses a nonpolar stationary phase?
   a. normal-phase chromatography
   b. reversed-phase chromatography
   c. ion-exchange chromatography
   d. size-exclusion chromatography
8. Which type of liquid chromatography uses a support with fixed charges to separate chemicals?
   a. reversed-phase chromatography
   b. ion-exchange chromatography
   c. size-exclusion chromatography
   d. affinity chromatography
9. Which of the following detectors might be used in LC to record an absorbance spectrum?
   a. photodiode array detector
   b. evaporative light scattering detector
   c. conductivity detector
   d. electrochemical detector
10. Separations in electrophoresis are based mainly on _____.
    a. interactions of analytes with a stationary phase
    b. the different volatilities of analytes
    c. the migration rates of analytes in an electric field
    d. the use of radioactive labels for detection
11. Which term is used to describe the pH at which a zwitterion has no net charge?
    a. electrophoretic mobility
    b. size or mass
    c. isoelectric point
    d. amino acid composition
12. Which of the following is NOT an advantage of capillary electrophoresis?
    a. It has low band-broadening due to effective removal of Joule heating.
    b. It can provide faster separations than gel electrophoresis.
    c. It can be used with large sample volumes.
    d. It can be used with an on-line detector.

Answers

1. b
2. b
3. c
4. c
5. d
6. b
7. b
8. b
9. a
10. c
11. c
12. c

Chapter 9

Electrochemistry

William Clarke[1] **and Paul D'Orazio**[2]

[1]*Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD, United States,* [2]*Instrumentation Laboratory, Bedford, MA, United States*

Learning objectives

After reviewing this chapter, the reader will be able to:

- List the major categories of electrochemical analysis.
- Describe the basis for ion-selective electrodes.
- Discuss the most widely used electrochemical methods in laboratory medicine.

Introduction

Electrochemistry is the area of chemistry that deals with the interconversion of chemical and electrical energy. Electrochemical methods of measurement used in clinical analysis can be divided into four major categories: potentiometry, amperometry (voltammetry), conductometry, and coulometry. In-depth fundamental reviews of these techniques can be found in textbooks dedicated to electrochemistry and electrochemical methods [1–3]. The routine use of electrochemical methods in the clinical laboratory can be traced to the introduction of modern blood gas analyzers in the 1950s [4]. During this time, Stow and Severinghaus developed the first practical sensor to measure the partial pressure of CO_2 in blood [5]; this discovery was followed closely by the development of the first pO_2 sensor by Clark to follow oxygenation of blood [6]. Electrochemical methods are advantageous because they can be adapted for sensor technologies to measure specific analytes rapidly, directly, and nondestructively in a complex sample matrix, such as whole blood. Because of this, there are several current clinical methods based on electrochemical analysis—a summary of these can be found in Table 9.1.

Potentiometric methods in clinical chemistry

Potentiometry is defined as the measurement of electrical potential (also designated electromotive force) between two electrodes when the cell current is zero. One electrode is typically designated as a reference electrode, while the second electrode is the indicator electrode (Fig. 9.1). The overall potential of a potentiometric cell is found by adding all the potential gradients that exist between different phases within the cell. It is possible to design a cell in such a way that all potential gradients (e.g., oxidation–reduction potential, membrane potential, diffusion potential) except for one are constant—the measured potential can then be related to the concentration of a specific analyte in the sample of interest. Potentiometry is by far the most utilized electrochemical method in clinical laboratories. Electrolytes and the partial pressure of carbon dioxide (pCO_2) can be measured directly in blood using potentiometric ion-selective electrodes (ISEs). At the core of the potentiometric cell shown in Fig. 9.1 is the silver–silver chloride (Ag/AgCl) electrode. The Ag/AgCl electrode is a specific type of potentiometric electrode known as a redox electrode, described briefly in the following section.

Redox electrodes

Redox potentials are the result of chemical equilibria involving transfer of electrons:

$$\text{Oxidized form(Ox)} + ne^- \leftrightarrow \text{Reduced form(Red)} \quad (9.i)$$

where n is the number of electrons involved in the reaction. The electrical potential generated by the redox couple is given by the Nernst equation:

$$E = E^o - 2.303\left(\frac{RT}{nF}\right) \times \log\left(\frac{a_{Red}}{a_{Ox}}\right) \quad (9.ii)$$

where E represents the half-cell potential, E^o the standard electrode potential when $a_{Red}/a_{Ox} = 1$, R the ideal gas constant (8.31 J K^{-1} mol^{-1}), T the absolute temperature (in Kelvin), n the number of electrons in the reaction, F the Faraday constant (96,487 coulomb/mol), and a the

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00009-0

activity. The term 2.303(R*T*/*n*F) equals 0.0592 V for a one electron transfer. The Ag/AgCl electrode is an example of a metal electrode that participates as a member of a redox couple. The Ag/AgCl electrode consists of a silver wire or rod coated with $AgCl_{(s)}$. This electrode is in contact with a solution of constant chloride activity, which sets the half-cell potential.

$$AgCl(s) + e - \leftrightarrow Ag^{\circ}(s) + Cl_{-} \qquad (9.iii)$$

and the equation governing the potential of the electrode is written as:

$$E = E^{o} - 0.0592 \times \log\left(\frac{a_{Ag} \times a_{Cl^-}}{a_{AgCl}}\right) \qquad (9.iv)$$

Since AgCl and Ag are solids, their activities are equal to 1, and the potential of the electrode is controlled by the activity of chloride ion contacting the electrode. The Ag/AgCl electrode is used both as an internal electrode in potentiometric ISEs and as a reference electrode half-cell of constant potential, required to complete a potentiometric cell (see Fig. 9.1).

TABLE 9.1 Clinical laboratory applications of electrochemical methods.

Methodology	Applications
Potentiometry	ISEs for pH, Na^+, K^+, Li^+, pCO_2 (internal element), Ca^{2+}, Mg^{2+}, and Cl^-
Amperometry	pO_2 electrode and H_2O_2 electrode (glucose biosensor internal element)
Conductometry	Measurement of hematocrit
Coulometry	Electrochemical titration of Cl^-

Ion-selective electrodes

ISEs are designed to respond selectively to one ionic species in a solution; a schematic of a potentiometric cell used to measure ion activity is shown in Fig. 9.1. The cell is composed of two half-cells: a reference electrode and an ISE. The ISE has a thin membrane that separates the sample from an internal electrolyte at constant activities for Cl^- and for the ion of interest. The potential for the entire cell (E) can be separated into several components (from Fig. 9.1):

- E_{ref} represents the potential at the internal Ag/AgCl element of the reference electrode.
- E_j represents the liquid junction potential at the interface of the concentrated electrolyte and sample.
- E_{int} represents the potential at the internal Ag/AgCl element of the ISE.
- E_{MEM} represents the potential developed across the ion-selective membrane.

The reference, internal, and liquid junction potentials can all be held constant, such that the only change in potential that occurs is across the ion-selective membrane (E_{MEM}), as a function of the activity of the ion of interest in the sample. Ion-selective membranes are typically composed of glass, crystalline, or polymeric materials, with either ion-exchange or ion-sequestering properties and high, but reversible, affinity for the ion of interest over other ions in the sample.

This electrochemical process can be described mathematically in the following manner—the potential generated across the ISE membrane consists of two potentials: one at the outer surface (E_{M1}) and one at the inner surface (E_{M2}). The membrane potential can be expressed as the difference between these two potentials:

$$E_{mem} = E_{M1} - E_{M2} \qquad (9.v)$$

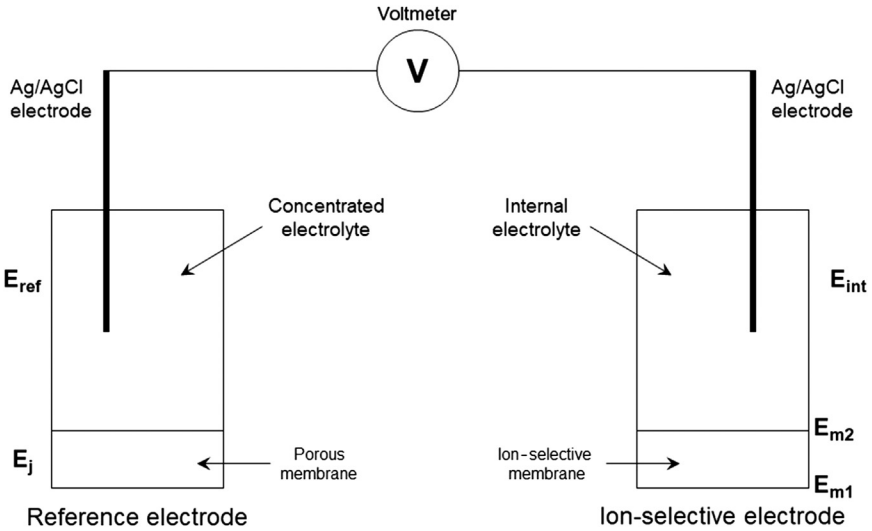


FIGURE 9.1 Schematic of a potentiometric cell using an ISE.

Both of the surface potentials are related to ion activities by Nernst-like relationships:

$$E_{M1} = j_1 - (0.0592/n) \times \log(a_1'/a_1) \quad (9.vi)$$

$$E_{M2} = j_2 - (0.0592/n) \times \log(a_2'/a_2) \quad (9.vii)$$

where j_1 and j_2 are constants, a_1 and a_2 are the activities for the ion of interest in solution, for the external and internal solutions with respect to the membrane, and n is the charge number for the ion of interest. The terms a_1' and a_2' are the ion activities *within* the membrane at the external and internal surfaces, respectively. If the two membrane surfaces have an equal number of ion-binding sites (which is typically the case), then j_1 and j_2 are identical, and so are a_1' and a_2'. By substituting Eqs. (9.vi) and (9.vii) into Eq. (9.v) and utilizing the fact that $j_1 = j_2$ and $a_1' = a_2'$, we arrive at the following equation:

$$E_{mem} = E_{M1} - E_{M2} = \left(\frac{0.0592}{n}\right) \times \log\left(\frac{a_1}{a_2}\right) \quad (9.viii)$$

This relationship shows that the membrane potential depends only on the ion activities of the solutions on either side of the membrane. However, recall that the ion activity for the internal solution is held constant, so the equation reduces to:

$$E_{mem} = E^{\circ} + \left(\frac{0.0592}{n}\right) \times \log a_1 \quad (9.ix)$$

where

$$E^{\circ} = \left(\frac{-0.0592}{n}\right) \times \log a_2 + E_{ref} + E_j + E_{int} \quad (9.x)$$

Therefore, when all other potentials in the cell are constant, the membrane potential is proportional to the ion activity in the external solution, or sample. It is important to note that this is not a potential generated directly by a redox reaction, but it is a phase boundary potential derived from transfer of the ion of interest across a concentration gradient—no oxidation or reduction reaction occurs. Sensitivity is determined by the change in potential corresponding to change in log units of activity for the ion of interest ($\Delta E/\Delta \log a$)—this corresponds to the Nernst slope (RT/nF). The theoretical value for the slope is 0.0592 V/concentration decade at 25°C for monovalent ions and 0.0296 for divalent ions.

TABLE 9.2 Ion selective electrodes in the clinical laboratory.

Electrode type	Clinical application
Glass membrane	pH (H^+) and Na^+
Ion exchange	K^+, Na^+, Cl^-, Ca^{2+}, Li^+, and Mg^{2+}
Gas sensing	pCO_2, pO_2, and pNH_3

The following are several important requirements for ISEs to be practical for the measurement of electrolytes in clinical specimens:

- Must have a Nernstian response to the ion of interest over the activity or concentration range of interest.
- Must have high selectivity for the ion of interest over other ions present in the sample.
- Must have a rapid response time.
- Must exhibit minimal drift over time.

The selectivity is dependent on analytical conditions; if the membrane has some affinity toward other ions present in the sample, then a complex diffusion potential arises that is not determined solely by the ion of interest. The Nicolsky equation [Interference section, Eq. (9.xxi)] can be used to describe the interference in terms of the activities of the ion of interest and the interfering ion.

There are three main categories of ISEs used today in clinical laboratories: glass membrane electrodes, ion-exchange or ion-sequestering electrodes based on polymer membranes, and gas-sensing electrodes; a list of clinically important electrolytes measured by ISEs can be found in Table 9.2.

Glass membrane electrodes

Glass electrodes are made from specially formulated melts of silicon dioxide with added oxides from various metals including aluminum and oxides of alkaline earth or alkali metal cations. Glass membranes range in thickness from 10 to 100 μm, and their electrical resistance is very high. Variation of the glass composition can change the selectivity of the membrane—glass electrodes for H^+, Na^+, Li^+, K^+, Rb^+, Cs^+, Ag^+, Tl^+, and NH_4^+ have been developed. However, glass electrodes for Na^+ and H^+ (pH) are the only ones with enough selectivity with respect to interfering ions to allow their use in the clinical laboratory.

Glass electrodes selective for H^+ (pH electrodes) were historically manufactured using silicon dioxide, sodium oxide, and calcium oxide in a molar ratio of 72.2:21.4:6.4 (Corning 015 glass). This glass has been shown to produce significant error for the measurement of pH when Na^+ is present at a physiological concentration and sample pH is more alkaline than 8, so other compositions have been tried and found to be more selective for H^+ ions. Presently, commercially available glass pH electrodes have sufficient selectivity for H^+ to allow error-free measurements between pH 7.0 and 8.0 in the presence of more than 0.1 mol/L Na^+. Glass electrodes can be made

selective for Na^+ by using a composition of silicon dioxide, sodium oxide, and aluminum oxide in the ratio of 71:11:18. These electrodes are insensitive to H^+ in the pH range of 6−10 [7]. Glass electrodes for pH and Na^+ are robust and require little maintenance—the main requirement being that the glass membrane requires occasional etching with a dilute solution of ammonium bifluoride to remove a stagnant layer of hydrated glass. This stagnant layer of glass can lead to a slowing of response time if allowed to build up over time.

Polymer membrane electrodes

Polymer membrane ISEs are used to measure a wide variety of clinically important ions directly in blood, including H^+, Na^+, K^+, Ca^{2+}, Cl^-, Li^+, and Mg^{2+}. There are two types of ISEs in this category: charged ion-exchanger type and ion-sequestering type based on neutral ion carriers, or ionophores. Both types typically use poly (vinyl chloride) (PVC) as a membrane matrix, an ion-exchanger or ionophore with selectivity for the ion of interest, a plasticizer to allow mobility of the ion-exchanger or ionophore within the membrane phase, and other additives in low concentration. An early ion-exchange-type ISE for Ca^{2+} used the calcium salt Ca^{2+}-bis(di-p-octylphenyl phosphate) in a PVC matrix [8]. An anion exchange-based electrode employing a lipophilic quaternary ammonium salt in PVC is still used for determination of Cl^- in blood. This type of electrode shows significant interference from ions in the sample which are more lipophilic in character than Cl^-, for example, thiocyanate and salicylate. A significant event in the development and application of polymer membrane ISEs occurred when Simon et al. discovered that the antibiotic valinomycin (Fig. 9.2) could be incorporated into a PVC membrane to yield a highly selective potassium electrode membrane [9,10]. Valinomycin is a neutral carrier (ionophore) that binds K^+ in the center of a ring of oxygen atoms. Other membranes of this type include those for NH_4^+, containing a mixture of the antibiotics nonactin and monactin, and several synthetic ionophores, for Ca^{2+}, Na^+, Li^+, and Mg^{2+} [11].

Gas-sensing electrodes

Gas electrodes are specially designed electrochemical cells for the measurement of specific gases in a gas mixture or in solution. The main examples of gas-sensing electrodes are those for carbon dioxide (CO_2) or ammonia (NH_3), both of which are based on a potentiometric pH electrode. (Electrochemical measurement of oxygen will be discussed in the next section.) Typically, a pH electrode is separated from the sample by a thin, permeable membrane (~20 μm) that only allows gases and water vapor to diffuse into the cell. The membranes can be made of polyethylene, Teflon, silicone rubber, or other microporous polymer materials. A diagram of a pCO_2 gas-sensing electrode can be seen in Fig. 9.3.

The carbon dioxide electrode is a self-contained electrochemical cell. The gas-permeable membrane is separated from the glass surface of a pH electrode by a thin layer of weakly buffered electrolyte solution, in this case a dilute sodium bicarbonate salt and a chloride salt. When placed in contact with the test solution, CO_2 diffuses through the membrane and rapidly dissolves into the electrolyte solution until equilibrium is reached. This process forms carbonic acid according to the chemical equation:

$$CO_2 + H_2O \Leftrightarrow H_2CO_3 \Leftrightarrow H^+ + HCO_3^- \qquad (9.xi)$$

and this shifts the pH of the internal electrolyte layer. The change in pH is a function of pCO_2 in solution and can be described by rearrangement of the Henderson−Hasselbalch equation:

$$pH = -s \times \log pCO_2 - \log a + pK' + \log cHCO_3^- \qquad (9.xii)$$

where s is the relative sensitivity of the electrode, a is the solubility coefficient of CO_2 (which varies with temperature), and K' is the first dissociation constant of carbonic acid. Therefore the change in pH is proportional to the change in log pCO_2. An ammonia electrode works the same way as the CO_2 electrode, except that the bicarbonate solution is replaced with an ammonium chloride solution. The measured pH of the ammonium chloride solution is linearly related to log pNH_3.

Ion-selective electrodes—units of measure and reporting for clinical application

Most analytical methods for electrolytes measure and report total ion concentration (c) usually expressed in millimoles per liter of sample (mmol/L). However, potentiometric measurements in undiluted blood, serum, or plasma (direct potentiometry) are sensitive to molality (m), millimoles of ion per mass of water in the sample (mmol/kg). The relationship between concentration and molality using the sodium ion as an example is:

$$cNa^+ = mNa^+ \times \rho H_2O \qquad (9.xiii)$$

where ρH_2O is the mass concentration of water in solution (kg/L). In aqueous solution, ρH_2O approximates 1 kg/L, while in normal plasma, ρH_2O is closer to 0.93 kg/L. In samples with elevated proteins or lipids, ρH_2O value may be as low as 0.80 kg/L. In these abnormal samples, the difference between concentration and molality may be as great as 20%, with concentration being the lower value. An advantage of direct potentiometry is that the technique is sensitive to molality and not affected by variations in the concentration of protein or lipids in the sample.

Methods that measure concentration, such as flame photometry, photometric methods, and potentiometric methods requiring sample dilution, report concentration and are affected by the level of proteins and lipids in the sample and may report falsely low electrolyte results (e.g., pseudohyponatremia) in cases of extremely elevated protein and lipid concentrations.

In addition to the difference between concentration and molality, direct potentiometry provides another unit of measurement known as activity (a). Potentiometric methods do not detect the concentration of complexed ions or ions that are "shielded" by other ions in solution (this can occur in solutions with high ionic strength). Ion activity is related to concentration with the following equation (using calcium as an example):

$$a\mathrm{Ca}^{2+} = \gamma\mathrm{Ca}^{2+} \times m\mathrm{Ca}^{2+} \tag{9.xiv}$$

where γ is the activity coefficient, and m is the molality of Ca^{2+} (moles Ca^{2+} per kg of water in the sample). The activity coefficient is primarily dependent on ionic strength of the solution, described by the Debye−Hückel equation:

$$\log \gamma = \frac{-\left(\mathrm{A} \times z^2 \times I^{1/2}\right)}{\left(1 + \mathrm{B} \times a \times I^{1/2}\right)} \tag{9.xv}$$

where A and B are temperature-dependent constants, z is the charge number of the ion, a is the ion size parameter for the specific ion, and I is the ionic strength of the solution.

$$I = \frac{1}{2}\sum m \times z^2 \tag{9.xvi}$$

While ionic activity is the appropriate parameter to discuss when considering biological equilibria and processes, most reference intervals and medical decision points for electrolytes are reported in molar concentration (or in equivalence units). This issue is addressed by using calibrator solutions with ionic strengths and compositions approximating normal human plasma. In this manner, the activity coefficient of each ion in the calibration solution is approximately the same as in the sample matrix, and the calibration can be performed in concentration units [12]. Unfortunately, the difference between molarity and molality cannot be calibrated out because water content [ρH_2O from Eq. (9.xiii)] in an aqueous calibrating

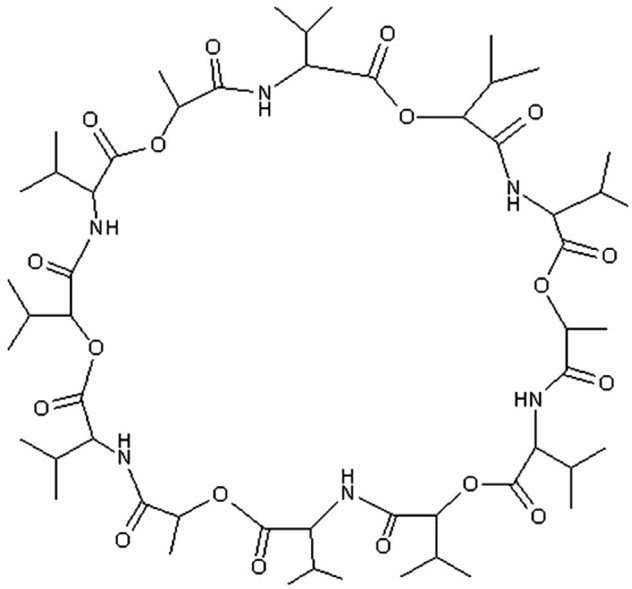


FIGURE 9.2 Chemical structure of valinomycin.

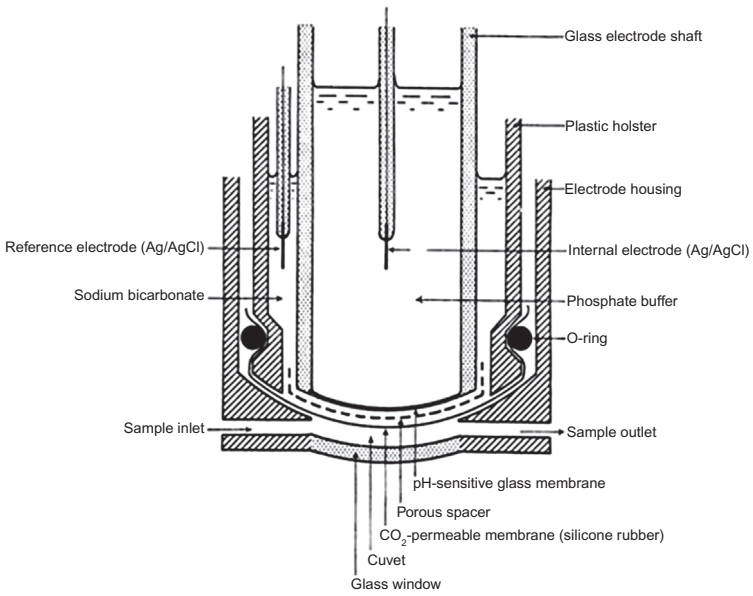


FIGURE 9.3 Schematic of a pCO_2 gas-sensing electrode.

solution (~0.99 kg/L) is much different than normal blood plasma (~0.93 kg/L). Direct potentiometry will report results approximately 6% greater than methods reporting concentration for normal samples because of this difference in water content between sample and calibrator (0.99/0.93 = 1.06). Most manufacturers of electrolyte measurement systems have addressed this problem in a practical way using correlation factors to standardize results from direct potentiometric measurements to units of concentration (mmol/L). However, when there is a significant deviation from typical water content of plasma (0.93 kg/L) due to hyperlipidemia or hyperproteinemia, the electrolyte concentration can be falsely reported as lower than the actual value (e.g., pseudohyponatremia) by indirect potentiometry. This is discussed in further detail in Chapter 36, Contemporary practice in clinical chemistry: blood gas and critical care testing.

Amperometric methods in clinical chemistry

Voltammetric measurements are used to study analytes based on the current−potential relationship when the electrochemical potential is varied. This approach takes advantage of the fact that certain molecules may be oxidized or reduced at an inert metal electrode when driven by an applied potential. When the redox reaction takes place, a current is generated in the cell; if the oxidation or reduction is limited by diffusion of the analyte to the electrode, the measured cell current is proportional to the amount of substance in the cell and will give a quantitative measurement of the analyte in solution. It is important to note that voltammetric methods measure concentration, not activity, and that the analyte of interest is consumed (at least partially) in the cell reaction.

Amperometry is a voltammetric technique where the applied potential is held constant, and the resultant current is used to quantitate the analyte of interest in solution. An amperometric cell typically consists of three electrodes: a *working* electrode, the *reference* electrode, and a third *counter (or auxiliary)* electrode. During amperometric measurements, the potential is measured and controlled between the working and reference electrodes, and the cell current is measured between the working and counter electrodes. This system avoids a possible error that may occur in a two-electrode system, as current passing between the working and reference electrodes could lead to a shift in the reference electrode potential, which would affect the applied cell potential during the measurement.

The most widely used application of amperometry in the clinical laboratory is for the determination of pO_2 in arterial and capillary blood. The Clark electrode [6] is a complete electrochemical cell consisting of a platinum working electrode (cathode) and a Ag/AgCl reference electrode (anode) (Fig. 9.4). In commercial instruments, the cathode is very small (~25 μm diameter) resulting in current levels in the picoampere (pA) range—when currents are this low, a counter electrode is generally not needed. The platinum cathode is covered by a thin film of electrolyte and separated from the test solution by a gas-permeable membrane. The cathode potential is adjusted to ~ −0.65 V relative to the anode, resulting in the following reactions:

$$\text{Cathode reaction: } O_2 + 2H_2O + 4e^- \rightarrow 4OH^- \quad (9.\text{xvii})$$

$$\text{Anode reaction: } Ag^+ + Cl^- \rightarrow AgCl + e^- \quad (9.\text{xviii})$$

Oxygen diffuses from the sample into the electrolyte layer, where it can be reduced at the platinum cathode. When no oxygen is present, the current is almost zero; when oxygen is present in the sample, the flow of current is directly proportional to the pO_2 in solution. It is important to remember that in order for this proportionality to hold true, the signal must be dependent on diffusion of oxygen to the cathode (diffusion-limited). The permeability and

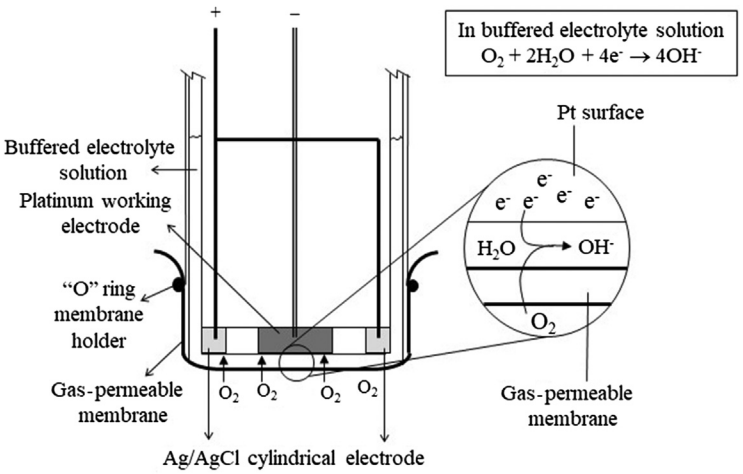


FIGURE 9.4 Diagram of a Clark pO_2 electrode.

thickness of the membrane are chosen to limit the diffusion of oxygen into the electrolyte layer and minimize variability in the diffusion coefficient of oxygen in the test solution. The membrane also functions as a barrier to keep proteins and other dissolved species from reaching the cathode and affecting the result.

Conductometric methods in clinical chemistry

Conductometric analyses are based on the ability of ions in solution to carry a current at a constant applied potential (conductivity). A potential is applied between two electrodes and then the current is measured; a decrease in resistance results in an increase in conductivity and thus an increased current occurs. The measured current is directly proportional to the sample conductivity.

Conductivity was first used to measure hematocrit in whole blood in the 1940s [13] and is still used today in many blood gas or critical care instruments. The basis for the measurement is that erythrocytes are natural electrical insulators because of their lipid-based membranes. As the proportion of red blood cells (RBCs) to plasma volume changes, so does the conductivity of the specimen. RBCs are less conductive than plasma, so as the volume fraction of RBCs increases, the conductivity decreases. This phenomenon can be described by Eq. (9.xix):

$$G_b = \frac{a}{1} + \left[\frac{H}{100} - H\right]c \qquad (9.xix)$$

where G_b represents the conductivity of whole blood, a the plasma conductivity, H the % hematocrit, and c the a factor for RBC orientation. In clinical practice, plasma conductivity must be corrected for effects from Na^+ and K^+ concentration in the sample; most clinical analyzers that measure hematocrit in this manner also measure these ions in conjunction with the whole blood conductivity, and correct for their effects prior to the hematocrit calculation. There are some significant limitations to hematocrit determination by conductivity. Dilution of clinical specimens with protein-free buffers, such as those used as plasma expanders during cardiopulmonary bypass surgery, lead to falsely low hematocrit values by conductivity, and pathologic conditions that lead to abnormally high serum protein levels may result in falsely high hematocrit values.

Coulometric methods in clinical chemistry

Coulometric methods consist of measuring the amount of charge (in coulombs) that passes between two electrodes in a fixed amount of time; the amount of charge is directly proportional to the oxidation or reduction of an analyte at one of the electrodes. Charge can be calculated from measured current because current is simply the amount of charge that passes per unit time (coulombs/second). The amount of charge can be related to the amount of electroactive substance (in moles) using Faraday's Law:

$$Q = z \times n \times F \qquad (9.xx)$$

where Q is the amount of charge passed during the reaction (coulombs), z is the number of electrons transferred in the redox reaction, n is the amount of analyte oxidized or reduced (in moles), and F is the Faraday constant.

Coulometry can be used in the clinical laboratory for the determination of chloride in serum (or plasma). Automatic coulometric titration instruments are available for this type of measurement. The instrument consists of a silver anode working electrode (silver is oxidized) and a platinum cathode reference electrode (hydrogen ion is reduced). A third indicator electrode is present (polarized at a negative potential) for detection of Ag^+ in solution—as Ag^+ is reduced by this electrode, a current is produced. As a constant current is applied to the working and reference electrodes, Ag^+ is produced, which combines with Cl^- in solution to form insoluble AgCl; as long as there are excess Cl^- ions, the Ag^+ concentration in solution will be low. When all the Cl^- in solution has been complexed, the Ag^+ in solution rapidly increases and produces a current at the indicator electrode (amperometry); when the current exceeds a set value, the titration is stopped. From the value of the constant applied current and the amount of time for the titration, charge can be calculated—from this value, the absolute number of Ag^+ ions produced may be calculated using Eq. (9.xx). Knowing that silver and chloride ions combine in a 1:1 stoichiometric ratio, the absolute number of Cl^- ions in the specimen can be determined. Thus, using the known volume of the original serum (or plasma) sample, it is possible to calculate the concentration of Cl^- in the sample. In the absence of interfering substances, coulometry is considered one of the most accurate electrochemical methods because the method measures the absolute amount of electroactive substance in the sample and is not dependent on steady-state or equilibrium processes. It is important to note that this is not a high-throughput technique, and it is typically used for sweat chloride measurement in the evaluation for cystic fibrosis. For general chemistry measurements in a high-throughput setting, chloride is typically measured by ISE.

Interferences for potentiometric and amperometric sensors

Potentiometric ISEs and amperometric sensors are the primary electrochemical methods used in clinical laboratory

applications. It is therefore appropriate to consider some specific types of interferences for these methods.

Some ISEs are subject to interference from other ions present in the sample. The response of an ISE in a solution consisting of the primary and interfering ion is described by the Nicolsky equation [a variation of the Nernst equation, Eq. (9.ix)].

$$E_{mem} = E^{\circ} + \left(\frac{0.0592}{n}\right) \times \log\left(a_i + K_{i/j} a_j^{zi/zj}\right) \quad (9.xxi)$$

where $K_{i/j}$ is the selectivity coefficient for the primary over interfering ion, determined experimentally [14], a_i and a_j are the activities of the primary and interfering ions, respectively, zi is the charge of primary ion, and zj is the charge of interfering ion. With knowledge of the selectivity coefficient and activity (or concentration) of the interfering ion, the concentration of the primary ion may be calculated from the measured cell potential.

Consider a practical example of interference with an ISE. The selectivity of an ISE for a primary ion (e.g., Ca^{2+}) over an interfering ion (e.g., Na^{+}) is given by the selectivity coefficient, expressed as $K_{Ca/Na}$ in this case. The selectivity coefficient is a quantitative measurement of the degree to which presence of an interfering ion will affect accuracy for measurement of the primary ion. Reciprocal of selectivity coefficient is a quantitative estimate of the extent to which the ISE will "favor" the primary ion over the interfering ion. Oesch et al. [15] have published an equation (derived from the Nicolsky equation) to calculate required selectivity coefficient for any ISE based on a tolerable level of interference due to presence of an interfering ion:

$$K_{i/j} = \frac{a_{i,\min}}{\left(a_{j,\max}\right)^{z_i/z_j}} \cdot \frac{p_{i/j}}{100} \quad (9.xxii)$$

where $K_{i/j}$ is the required value of the selectivity coefficient, $a_{i,\min}$ is the lowest expected activity of the primary ion, $a_{j,\max}$ is the highest expected activity of the interfering ion, $p_{i/j}$ is the highest tolerable error for measurement of the primary ion due to interference (in %), and z_i and z_j are charges of the primary and interfering ions, respectively

For example, measurement of Ca^{2+} at a concentration of 1.10 mmol/L in the presence of 160 mmol/L Na^{+}, with a maximum tolerable error of 1%, requires a sensor with $K_{Ca/Na} = 4.3 \times 10^{-4}$, which is equivalent to a preference for Ca^{2+} over Na^{+} of 2325:1. Older Ca^{2+} sensor designs utilizing charged calcium ion-exchangers, such as didecylphosphoric acid or related compounds, exhibit a preference for Ca^{2+} over Na^{+} no better than 200:1 and do not meet the above selectivity requirements [16]. Newer Ca^{2+} sensors, based on neutral carriers for calcium ion, generally meet the selectivity requirements with preference for Ca^{2+} over Na^{+} of 3000:1 or better [17]. However, even these sensors may be challenged in their selectivity over Na^{+} at very low Ca^{2+} concentrations, depending on other additives in the Ca^{2+} sensing membrane and other sources of imprecision or bias in the analyzer, such as surfactants present in calibrating, rinse, or quality control solutions which can produce deterioration in selectivity [18].

Potentiometric and amperometric gas electrodes are particularly sensitive to temperature and barometric pressure (BP) in the lab. When gas electrodes are calibrated, the partial pressure of the gas is dependent on the BP, percentage of gas in the mixture, and the partial pressure of water vapor as described by the following equation:

$$pX = (\mathrm{BP} - 47)\left(\frac{\%\ \text{of gas } X}{100}\right) \quad (9.xxiii)$$

where BP is the ambient barometric pressure in mm Hg, and 47 is the partial pressure of water vapor at 37°C. It is obvious from the equation that changes in humidity or temperature can significantly affect the calibration of the instrument and lead to erroneous results. Temperature in the analyzer must be controlled to tenths of degree celsius in order to meet precision requirements for blood gas analyzers. In addition, the inherent drift of electrochemical gas sensors means that frequent calibration is required—2-point calibrations every 2−4 hours and a 1-point calibration every 15−20 minutes. Most commercial blood gas analyzers are programmed to automatically perform these calibrations during the day at predetermined intervals. Because gas electrodes for pO_2 and pCO_2 are isolated from the sample by gas-permeable membranes, they are largely free of chemical interfering substances from the sample matrix. Early generation pO_2 sensors were reported to show interference from anesthetic gases such as nitrous oxide present in the sample [4]. These substances are electrochemically active and have the ability to diffuse across the gas-permeable membranes used in early pO_2 sensors. However, newer polymeric membranes found in later generations of blood gas electrodes have mostly eliminated this interference.

Biosensors based on electrochemical methods

A biosensor is a specific type of sensor consisting of a biological recognition element and a chemical transducer. The recognition element interacts selectively with an analyte of interest in the sample. This interaction results in a measurable change in a solution property, such as formation of a product or consumption of a reactant. The chemical transducer converts the solution property into a quantifiable electrical signal. Using enzymes as

recognition elements in combination with potentiometric or amperometric sensors has allowed the development of practical electrochemical sensors for analytes, which are electrochemically inactive. Biosensors for the measurement of analytes such as glucose, lactate, urea and creatinine directly in blood are used today in combination with sensors for blood gases and electrolytes in systems for critical care testing.

Enzyme-based biosensors with amperometric and potentiometric detection

Enzyme-based biosensors using electrochemical transducers, specifically amperometric electrodes, are the biosensors most commonly used for clinical analyses. The majority of today's blood glucose meters use enzyme-based biosensors with amperometric detection, with glucose sensors produced in single-use formats. Clark and Lyons developed the first amperometric biosensor; it was used to measure glucose in blood and was based on immobilizing glucose oxidase on the surface of an amperometric pO_2 sensor [6] (see diagram of pO_2 sensor in Fig. 9.4). A solution of glucose oxidase was physically entrapped between the gas-permeable membrane of the pO_2 electrode and an outer semipermeable membrane. The outer membrane was of a low-molecular-weight cutoff to allow substrate (glucose) and oxygen from the sample to pass, but not proteins and other macromolecules. In this way, the enzyme could be concentrated at the sensor's surface. Oxidation of glucose, catalyzed by glucose oxidase as follows:

$$\text{Glucose} + O_2 \xrightarrow{\text{glucose oxidase}} \text{gluconic acid} + H_2O_2 \tag{9.xxiv}$$

consumes oxygen near the surface of the sensor. The rate of decrease in pO_2 is a function of the glucose concentration and is monitored by the pO_2 electrode. A steady-state reduced partial pressure of oxygen can be achieved at the surface in a short period of time, yielding a steady-state current value that is inversely proportional to glucose concentration in the sample. If the polarizing voltage of the pO_2 electrode is reversed, making the platinum electrode positive (anode) relative to the Ag/AgCl reference electrode, and if the gas-permeable membrane is replaced with a hydrophilic membrane containing immobilized glucose oxidase enzyme, it is possible to oxidize the H_2O_2 produced in Eq. (9.xxiv) as follows:

$$H_2O_2 \xrightarrow{\text{applied voltage}} 2H^+ + O_2 + 2e^- \tag{9.xxv}$$

In practice, a sufficiently high voltage must be applied to the platinum anode to drive the oxidation of the hydrogen peroxide. An applied voltage of +0.7 V or greater (relative to Ag/AgCl) is typically used. In this case, the steady-state current produced is directly proportional to the concentration of glucose in the sample. Substitution of other oxidase enzymes in Eq. (9.xxiv) allows practical biosensors for other analytes of clinical interest such as lactate [19] and creatinine [20].

Substances found in blood, which are oxidizable at the applied potential shown in Eq. (9.xxv), are potential interfering substances for amperometric biosensors, particularly amperometric biosensors based on oxidation of hydrogen peroxide. Direct oxidation of the interfering substance at the anode produces the interfering signal. Such substances include ascorbic acid, uric acid, acetaminophen, and other therapeutic drugs [21]. Elimination of interfering substances from altering biosensor response often requires an additional membrane between the enzyme and the electrode surface. A low-molecular-weight cutoff membrane such as cellulose acetate may be used to prevent diffusion of the interfering substance to the electrode surface while maintaining permeability toward hydrogen peroxide. A combination of other techniques such as a second correcting electrode without enzyme to compensate for the signal from oxidation of interferences, the incorporation of anionic sites in the outer membrane to reject negatively charged interfering substances, and the use of lower applied potentials have all been successfully used to reduce the effect of interfering substances on amperometric biosensor response.

ISEs have been used as transducers for potentiometric biosensors. An example is the commercially available biosensor for urea based on an ISE for ammonium ion [22]. The enzyme urease is immobilized at the surface of the ISE and catalyzes the hydrolysis of urea to NH_3 and CO_2.

$$\text{Urea} \xrightarrow{\text{urease}} 2NH_3 + CO_2 \tag{9.xxvi}$$

Ammonia is hydrolyzed to form NH_4^+, which is sensed by the ammonium ISE. The ammonium ISE is based on the ionophore nonactin and suffers interference from Na^+ and K^+. These ions are typically measured simultaneously with urea and used to correct the output of the urea sensor using the Nicolsky equation [see Eq. (9.xxi)]. The selectivity coefficients for NH_4^+ over K^+ and Na^+ must be known or determined as part of the calibration process.

Enzyme-based electrochemical biosensors require another key element to assure practical use in blood. An outer membrane is commonly used to restrict diffusion of substrate to the enzyme layer, avoiding saturation of the enzyme and providing extended linearity toward the analyte of interest at substrate concentrations substantially higher than the K_m of the enzyme. This membrane also mechanically protects the enzyme layer and screens out macromolecules and RBCs from contact with the enzyme.

Track-etched polycarbonate membranes are commonly used. Hydrophobic coatings with lipids or silanes have been used to improve the biocompatibility of the polycarbonate membrane surface. Membranes based on polyvinylchloride, polyurethane, and silicone emulsions have also been used.

Affinity biosensors with electrochemical detection

Affinity sensors are a special class of biosensors in which the immobilized biological recognition element is a binding protein, an antibody (immunosensors), or an oligonucleotide (DNA sensors) that have high binding affinity and high specificity toward a clinically important analyte. Such sensors have been developed as alternatives to conventional binding assays to enhance the speed and convenience of a wide range of assays that would be typically run on sophisticated immunoassay analyzers. Ideally, direct binding of the immobilized species to its target in a clinical sample should yield a sensor signal proportional to the concentration of the analyte. However, "direct" sensing affinity biosensors at analyte concentrations that would cover the full range of clinical applications are very difficult to achieve. Further, high affinity of such binding reactions, required to achieve optimal sensitivity, limits the reversibility of such devices. Affinity sensors-based electrochemical transduction are typically single-use devices, thus obviating the need for some type of regeneration step (pH change, etc.) to dissociate the tight binding between the recognition element and the target.

The number of research reports related to affinity biosensors continues to increase. However, commercialization has lagged behind research output, and movement of these types of biosensors from the research laboratory to the clinical laboratory has been slow. Affinity biosensors with promise of clinical utility are typically based on labeled reagents such as enzymes and electrochemical tags; hence, they function more like traditional binding immunoassays, except that one recognition element is immobilized on the surface of an electrochemical transducer. For example, electrochemical pO_2 sensors have been employed to perform heterogeneous enzyme immunoassays (sandwich or competitive type), using catalase as a labeling enzyme (catalyzes H_2O_2 $2\,H^+$, and O_2). Capture antibodies are immobilized on the outer surface of the gas-permeable membrane. After binding equilibration between antigen and capture antibody, followed by a washing step, a second, catalase-labeled reporter antibody interacts with surface-bound antigen. The activity of bound enzyme is detected by adding the substrate (H_2O_2) and following the increase in current generation caused by local production of oxygen near the surface of the sensor. Some commercial examples of immunosensors do exist, primarily in the unit-use, disposable format designed for point-of-care testing. One successful commercial example is a cartridge-based device for cardiac troponin I on the i-STAT handheld analyzer (Abbott Point of Care) [23]. This sensor uses a sandwich immunoassay format with electrochemical detection. Capture antibody is immobilized on a gold electrode for recognition and capture of troponin in the blood sample. A second, reporter antibody is labeled with alkaline phosphatase and binds with surface-captured troponin antigen. Following incubation and washing steps, the substrate p-aminophenyl phosphate is introduced and product of the enzymatic reaction (p-aminophenol) is detected amperometrically by oxidation at the gold electrode. Magnitude of the oxidative current is proportional to the concentration of troponin in the sample.

Affinity biosensors based on oligonucleotide binding (DNA sensors or "gene" sensors) in which a segment of DNA complementary to a target strand is immobilized on a suitable electrochemical sensor have been demonstrated. These devices operate in either direct (based on electrochemical oxidation of guanine in target DNA) or indirect (with exogenous electrochemical markers/labels) transduction modes. Although most of the proposed electrochemical DNA biosensors require amplification methods, such as PCR, to multiply small amounts of DNA into measurable quantities, some are sensitive enough to eliminate the need for target amplification. Nanotechnology has been proposed, in an indirect format, for signal amplification. For example, a capture probe DNA is immobilized on a gold electrode. Reporter probes with electrostatically bound ruthenium complexes $[Ru(NH_3)_6]^{3+}$ are loaded onto gold nanoparticles (AuNP) and are capable of hybridizing with one of two sequences on target DNA. The other sequence on the target DNA is capable of hybridizing with the immobilized capture probe. Hybridization events on the electrode surface bring multiple reporter probes for each AuNP. Electroactive $[Ru(NH_3)_6]^{3+}$ is reduced at the electrode surface, and the coulometric signal is proportional to the concentration of target DNA [24]. A commercial example of electrochemical DNA sensing, along with AuNP probes without need for PCR amplification, is the Verigene system (Luminex), capable of detecting single-nucleotide polymorphisms related to common genetic disorders, such as thrombophilia, alterations of folate metabolism, cystic fibrosis, and hemochromatosis.

Another example of an electrochemical "gene" sensor array uses electrochemical probes that are selectively inserted into hybridized DNA duplexes. In one approach, after the immobilized capture of oligo anchored to the electrode surface is allowed to bind a target sequence, hybridization is detected by exposing the surface of the

electrode to an exogenous electroactive species (Co[III] tris-phenanthroline, ruthenium complexes, etc.) that intercalates within the duplex, but not to single-stranded DNA. After unbound electroactive species are removed by washing, the presence of hybridization is readily detected by voltammetry, scanning the potential of the underlying electrode to oxidize or reduce any intercalated electroactive species, with the measured current being proportional to the number of duplex DNA species on the surface of the electrode.

References

[1] A.J. Bard, L.R. Faulkner, Electrochemical Methods: Fundamentals and Applications, second ed, John Wiley, New York, 2001, p. xxi.

[2] D.C. Harris, second ed., Quantitative Chemical Analysis, 1v, W. H. Freeman and Co., New York, 1987, pp. 297–383.

[3] D.A. Skoog, J.J. Leary, fourth ed., Principles of Instrumental Analysis, xii, Saunders College Pub., Fort Worth, TX, 1992, pp. 462–488.

[4] C.E. Hahn, Electrochemical analysis of clinical blood-gases, gases and vapours, Analyst 123 (1998) 57R–86R.

[5] J.W. Severinghaus, A.F. Bradley, Electrodes for blood pO2 and pCO2 determination, J. Appl. Physiol. 13 (1958) 515–520.

[6] L.C. Clark Jr., C. Lyons, Electrode systems for continuous monitoring in cardiovascular surgery, Ann. N. Y. Acad. Sci. 102 (1962) 29–45.

[7] G. Eisenman, D.O. Rudin, J.U. Casby, Glass electrode for measuring sodium ion, Science 126 (1957) 831–834.

[8] J.W. Ross, Calcium-selective electrode with liquid ion exchanger, Science 156 (1967) 1378–1379.

[9] B.T. Kilbourn, J.D. Dunitz, L.A. Pioda, W. Simon, Structure of the K + complex with nonactin, a macrotetrolide antibiotic possessing highly specific K + transport properties, J. Mol. Biol. 30 (1967) 559–563.

[10] L.A. Pioda, W. Simon, H.R. Bosshard, C.H. Curtius, Determination of potassium ion concentration in serum using a highly selective liquid-membrane electrode, Clin. Chim. Acta 29 (1970) 289–293.

[11] W. Simon, E. Pretsch, W.E. Morf, D. Ammann, U. Oesch, O. Dinten, Design and application of neutral-carrier based ion-selective electrodes, Analyst 109 (1984) 207–209.

[12] H.F. Osswald, H.R. Wuhrmann, Calibration standards for multi ion analysis in whole blood samples, in: D.W. Lubbers, H. Acker, R.P. Buck, et al. (Eds.), Progress in Enzyme and Ion Selective Electrodes, Springer Verlag, Berlin, 1981, pp. 74–78.

[13] R.L. Rosenthal, C.W. Tobias, Measurement of the electrical resistance of human blood: use in coagulation studies and cell volume determinations, J. Lab. Clin. Med. 33 (1948) 1110–1122.

[14] R.P. Buck, E. Lindner, Recommendations for nomenclature of ion selective electrodes, Pure Appl. Chem. 66 (1994) 2527–2536.

[15] U. Oesch, D. Ammann, W. Simon, Ion-selective membrane electrodes for clinical use, Clin. Chem. 32 (1986) 1448–1459.

[16] G.J. Moody, R.B. Oke, J.D.R. Thomas, A calcium-sensitive electrode based on a liquid ion exchanger in a poly(vinyl chloride) matrix, Analyst 95 (1970) 910–918.

[17] P. Anker, E. Wieland, D. Ammann, R.E. Dohner, R. Asper, W. Simon, Neutral carrier based ion-selective electrode for the determination of total calcium in blood serum, Anal. Chem. 53 (1981) 1970–1974.

[18] E. Malinowska, M.E. Meyerhoff, Influence of nonionic surfactants on the potentiometric response of ion-selective polymeric membrane electrodes designed for blood electrolyte measurements, Anal. Chem. 70 (1998) 1477–1488.

[19] J. Aduen, W.K. Bernstein, T. Khastigir, et al., The use and clinical importance of a substrate specific electrode for rapid determination of blood lactate concentration, JAMA 272 (1994) 1678–1685.

[20] A.J. Killard, M.R. Smyth, Creatinine biosensors: principles and designs, Trends Biotechnol. 18 (2000) 433–437.

[21] G.J. Kost, T.H. Nguyen, Z. Tang, Whole blood glucose and lactate: trilayer biosensors, drug interference, metabolism and practice guidelines, Arch. Pathol. Lab. Med. 124 (2000) 1128–1134.

[22] G.G. Guilbault, Enzyme electrode probes, in: G.G. Guilbault (Ed.), Analytical Uses of Immobilized Enzymes, Marcel Dekker, New York, 1984, pp. 112–244.

[23] F.S. Apple, R. Ler, A.Y. Chung, et al., Point-of-care i-STAT cardiac troponin I for assessment of patients with symptoms suggestive of acute coronary syndrome, Clin. Chem. 52 (2006) 322–325.

[24] J. Zhang, S.P. Song, L.H. Wang, et al., A gold nanoparticle-based chronocoulometric DNA sensor for amplified detection of DNA, Nat. Protoc. 2 (2007) 2888–2895.

Self-assessment questions

1. Which of the following describes the cathode in an electrochemical cell?
   a. loss of electrons occurs; reduction
   b. gain of electrons occurs; reduction
   c. loss of electrons occurs, oxidation
   d. gain of electrons occurs; oxidation
2. Potentiometric analysis involves measurement of:
   a. potential while as current is varied
   b. conductance of material under an applied potential
   c. magnitude of current at a fixed potential
   d. electromotive force while current is zero
3. Which of the following is routinely measured using conductometry?
   a. hematocrit
   b. lipid content of blood
   c. sodium ion concentration
   d. bone density
4. Potassium can be measured using which of the following techniques?
   a. glass membrane ion-selective electrode
   b. coulometric titration
   c. gentamicin-based ion-exchange electrode
   d. none of the above
5. The Nernst equation is used to:
   a. determine current under nonstandard conditions
   b. determine variance from standard potentials under nonstandard conditions
   c. calculate activity coefficient from the ionic strength of solution
   d. calculate hematocrit from conductance measurements
6. Which of the following can be a potential source of error for potentiometric measurements in plasma?
   a. hyperlipidemia
   b. multiple myeloma
   c. dilution of specimen with water
   d. all of the above
7. In potentiometric gas-sensing electrodes, measurement of the partial pressure of the gas is described by:
   a. partial pressure is directly proportional to pH
   b. partial pressure is directly proportional to measured current
   c. pH is directly proportional to the log of partial pressure
   d. partial pressure is directly proportional to the log of pH
8. Successive measurements with a pCO_2 electrode will be continuously lower due to consumption of the analyte during the analytical process. True or False?
   a. True
   b. False

Answers

1. b
2. d
3. a
4. d
5. b
6. d
7. c
8. b

Chapter 10

Mass spectrometry

Stefani N. Thomas

Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, MN, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- List the basic components of a mass spectrometer.
- Compare and contrast the various methods for ionization and mass analysis.
- Explain the concept of tandem mass spectrometry.
- Summarize the advantages and limitations of mass spectrometry for use in the clinical laboratory.

In 1910 Thompson [1] built the first instrument used to measure mass-to-charge (*m/z*) values of gaseous ionized atoms: a mass spectrometer. Soon thereafter, mass spectrometry (MS) became essential for the study of various molecules, and from 1913 to 1945 MS was predominantly used for the analysis of isotopes and atomic weights [2]. The applications of MS have since expanded to include a myriad of qualitative and quantitative applications in many fields, including forensics, basic and translational research, environmental science, pharmaceutical science, geology, astronomy, and, of course, laboratory medicine.

Initial applications of MS were relegated to highly specialized laboratories staffed with individuals with considerable technical expertise due to the knowledge base required to effectively operate the complex MS instrumentation. However, recent advances in MS instrumentation have occurred concomitantly with relatively decreased levels of complexity from the standpoint of instrument operation. Hence, MS now is among the routinely used cadre of instrumentation in many clinical laboratories. Applications of MS in the clinical laboratory include therapeutic drug monitoring, testing for inborn errors of metabolism, steroid analysis, and elemental analysis. Indeed, MS is gradually transforming the practice of laboratory medicine, largely due to its analytical sensitivity and specificity [3].

Understanding the fundamental tenets of this powerful technology is a requisite step toward the effective application of MS, particularly in the clinical laboratory. This chapter addresses the basic concepts and definitions of MS with a particular emphasis on instrumentation. Brief mention is made of some of the clinical applications of MS; for a more thorough review of this topic, the readers are encouraged to refer to Chapter 21, Applications of mass spectrometry in the clinical laboratory.

Basic mass spectrometry concepts

MS is a powerful analytical technique that can be used to identify and quantify analytes using the *m/z* of ions generated from a sample. Small molecules, peptides, and proteins are among the clinically relevant analytes that are routinely analyzed by MS. The analysis of such analytes by MS can be performed using systems that exist in several different instrument configurations; however, most MS systems comprise five main components: (1) an ionization source; (2) a vacuum system; (3) a mass analyzer; (4) a detector; and (5) a processor (Fig. 10.1).

MS analysis requires an initial ionization step in which an ion, defined as an atom or molecule with a net electric charge due to the loss or gain of one or more electrons, is produced from a neutral atom or molecule. These ions must be present in the gas phase prior to mass analysis.

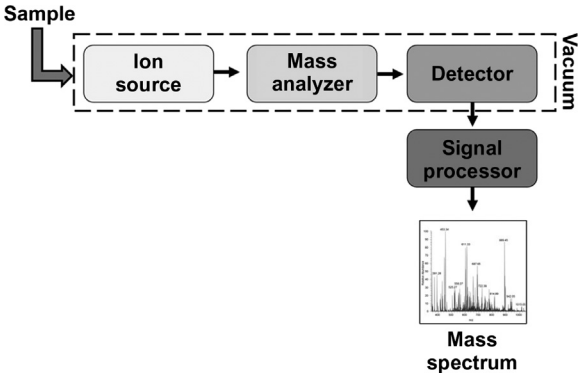


FIGURE 10.1 Overview of the components of a mass spectrometry instrument.

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00010-7

Ion separation requires that the ions do not collide with other molecules, as they traverse the ion path inside of the mass spectrometer, which occurs under vacuum. The magnitude of the vacuum pressure (10^{-3}–10^{-9} torr) varies based on the type of mass analyzer. Whereas Fourier-transform ion cyclotron resonance (FTICR) analyzers require a low pressure of $\sim 10^{-9}$ torr, ion trap mass analyzers operate under higher pressure of 10^{-3}–10^{-5} torr. Vacuum pumps with higher pumping capacities are generally associated with lower detection limits. Mass analyzers separate ions based on their *m/z* ratio prior to their delivery to the detector. The function of the detector is to sense ions, as they travel from the mass analyzer, to increase the signal from the ions (often by several orders of magnitude), and then to convert the charge from the ions to a current. Finally, the signal processor converts current into a peak on the mass spectrum. The "Detectors and processors" section below addresses these concepts in further detail.

A mass spectrum represents the relative abundance or signal intensity of each ion plotted as a function of *m/z* (Fig. 10.2). The accurate interpretation of mass spectra requires an understanding of the concept of molecular mass (also known as molecular weight), which is measured in units of daltons, where 1 dalton (Da) is equivalent to 1/12 of the atomic mass of the most abundant isotope of a carbon atom in its lowest energy state (^{12}C). Small molecules (<1000 Da) typically only carry one charge; thus the *m/z* value is identical to the mass of the molecular ion. However, larger molecules such as peptides or proteins often carry multiple ionic charges, which translates to the *z* value being an integer >1 and the *m/z* value being less than the molecular mass of the analyte.

In a mass spectrum, the peak representing the intact ion from the original molecule is referred to as the molecular ion peak. The ion present in the greatest amount is termed the base peak and is assigned a relative value of 100%. The ability of a mass spectrometer to separate nearby masses from each other is known as resolution, which is defined as $(m/z)/(\Delta m/z)$, where $\Delta m/z$ is the width of the mass spectral peak that is typically measured at 50% of the peak height and is referred to as the full width half height (FWHH) or full width half maximum (FWHM) resolution. FWHH and FWHM are measures of resolution based on a single peak. A definition of resolution that applies to two adjacent, symmetric peaks is based on the 10% valley, which defines $\Delta m/z$ as the distance between two peaks of equal intensity that are separated, such that the valley between the peaks is 10% of the peak height. Another important parameter when interpreting mass spectra is mass accuracy, which refers to how close a measured mass is to its expected (true) mass. Mass accuracy is defined as the difference between the measured and expected masses divided by the true mass. Resolution and mass accuracy vary greatly among different types of MS systems.

In addition to the detection of intact ions, mass spectrometers can also be used for the detection of fragment ions that are generated as a result of energy that is imparted into the ionized analyte, causing internal bonds to break and resulting in fragmentation. The fragment ions can then be used for the purpose of the structural determination of the intact ion. Fragmentation can occur in the ion source as well as in a separate dissociation, or collision, cell inside a tandem mass spectrometer. Tandem mass spectrometers will be addressed in a subsequent section of this chapter.

Any discussion of mass spectra interpretation warrants mention of the influence of isotopes. Chemical elements can be composed of a single or multiple isotopes. In the context of MS, isotopes can reveal the charge state of a peak, the types of elements that could be contained in a particular molecule, and they can indicate the elemental composition of a molecule via accurate mass determination. The monoisotopic mass in a mass spectrum refers to the mass of the isotopic peak whose elemental

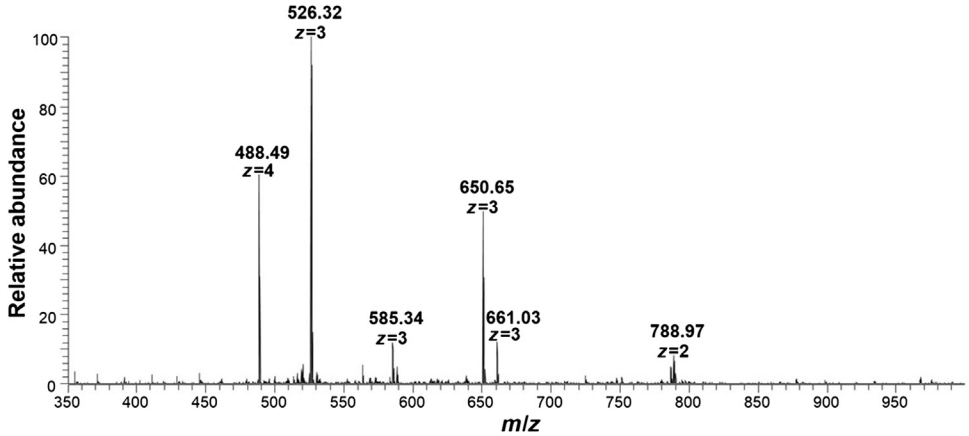


FIGURE 10.2 Representative mass spectrum of complex peptide mixture.

composition comprises the most abundant isotope of the constituent element, whereas the average mass is the average of the isotopic masses weighted by the isotopic abundances. Among the many strengths of MS as an analytical method is the ability to distinguish between ions with the same chemical formula that have different masses due to the difference in their isotopic composition. A comprehensive discussion of the interpretation of mass spectra is provided by McLafferty and Tureek [4].

Sample introduction

Sample preparation is essential for MS analyses that yield useful and reliable data. Given that mass spectrometers are considered close to being universal detectors based on their ability to detect a broad array of ionizable molecules, particular attention must be given to the physical and chemical properties of the analyte of interest in the context of its background matrix when determining the specific protocol for sample preparation prior to MS analysis. Clinical specimens such as plasma, serum, and urine contain background analytes with a wide dynamic range of abundance, which can complicate MS analysis if suboptimal sample preparation is performed.

Ion suppression or enhancement can occur due to the presence of sample constituents such as salts, phospholipids, surfactants, and ion-pairing agents that suppress or enhance ionization of the analyte of interest, which can significantly confound the interpretation of MS data. For example, in applications entailing liquid chromatography coupled with MS, nonvolatile or less volatile components found in a sample can reduce droplet formation and the efficiency of solvent evaporation in the ionization source, resulting in reduced ion formation. Ionization efficiency is a measure of how well gas-phase ions are formed from analyte molecules and it is impacted by the analyte itself, the concentration of the analyte, and the type of ionization process used. Several techniques for determining the degree of ion suppression and enhancement have been described [5,6].

Among the common sample preparation methods that are undertaken for clinical MS analysis are protein precipitation followed by centrifugation, solid-phase extraction, liquid−liquid extraction, derivatization, and affinity enrichment. The buffer into which the sample is extracted should be compatible with the downstream chromatographic and ionization conditions.

Sample complexity must be taken into consideration when determining the method by which a sample is to be introduced into a mass spectrometer. Liquid and gas chromatographs are versatile analytical instruments that can be readily coupled with mass spectrometers. The resolving capabilities of these instruments combined with the specificity of MS have rendered liquid chromatography-mass spectrometry (LC-MS) and gas chromatography-mass spectrometry (GC-MS) powerful analytical techniques for clinical analysis. Broad spectrum molecular weight polar and nonpolar compounds that can be readily dissolved in liquid, such as proteins, vitamins, and certain drugs, can be analyzed using LC-MS, whereas GC-MS analysis is particularly suitable for the analysis of low molecular weight and nonpolar or volatile analytes such as organic acids, cholesterol, steroid hormones, and xenobiotic compounds [7].

Ion sources

Regardless of the method of sample introduction to the mass spectrometer, the generation of gas-phase ions is required for MS detection. When gas-phase molecules are introduced directly into an ionization source from a gas chromatography (GC) system, electron ionization (EI) and chemical ionization (CI) are the most commonly used techniques. When a liquid chromatography (LC) system is coupled with a mass spectrometer for the introduction of liquid-state analytes, ionization under atmospheric or near-atmospheric conditions is conducted using methods such as electrospray ionization (ESI), atmospheric pressure chemical ionization (APCI), or atmospheric pressure photoionization (APPI). Other commonly used ionization techniques include inductively coupled plasma (ICP), matrix-assisted laser desorption ionization (MALDI), desorption electrospray ionization (DESI), rapid evaporative ionization (REI), and thermal desorption (TD).

Electron ionization

EI is the most common form of sample ionization used for GC-MS applications. It occurs in a gas-phase reaction described by the following equation:

$$M + e^- \rightarrow M^{+\cdot} + 2e^- \qquad (10.i)$$

where M is the neutral molecule undergoing ionization, e^- is the electron, and M^+ is the resulting charged radical cation. Uncharged molecules in the gas phase are bombarded with a stream of electrons that are accelerated through an electric field at an ion energy of 70 eV. The potential difference energizes the electron source, such that collision with neutral molecules results in the ejection of two electrons and the formation of a charged radical cation.

This process frequently causes numerous cleavage reactions that give rise to fragment ions, which can convey structural information about the analyte of interest. Consequently, EI is often referred to as a "hard ionization" method due to the extensive fragmentation that occurs. The fragment ions can then be compared with

curated libraries of established EI fragmentation patterns to aid in structural identification.

Chemical ionization

CI is another ionization method that is commonly used for GC-MS applications; however, it is considered a "soft ionization" method due to the formation of primarily molecular ions. In CI, a proton is transferred to, or abstracted from, a gas-phase molecule such as methane, ammonia, or isobutane. The transfer results in protonated ions that carry a net positive charge (MH^+). An example of this reaction using methane as the CI reagent gas is as follows:

$$CH_5^+ + M \rightarrow CH_4 + MH^+ \qquad (10.ii)$$

Negative ions can also be produced by CI wherein protons are extracted from basic gases. These negative ions are formed when thermalized electrons are captured by electronegative functional groups, such as fluorine or chlorine atoms, within the molecule of interest. Negative ion CI is commonly used for the quantification of benzodiazepines and other drugs.

Electrospray ionization

The development of ESI greatly facilitated the broadening of the applications to which MS can be applied. ESI enables the delivery of ions from liquid solutions into the gas phase, including ions from large nonvolatile molecules such as proteins. A solution containing ionic analyte molecules dissolved in a volatile liquid is pumped through a charged capillary to which a 1−5 kV voltage is applied, and it is emitted as an aerosol comprising small droplets, ~10 μm in diameter. The aerosol is thought to be produced by a process involving the formation of a Taylor cone. A gas such as nitrogen facilitates the nebulization of the liquid and the evaporation of the solvent in the droplet. As the solvent evaporates, the analyte molecules are drawn closer together, repel each other, and cause the droplets to break into smaller droplets in a process known as Coulombic fission. This is an iterative process that results in the formation of desolvated analytes, which are then drawn into the orifice of the mass spectrometer. The ESI process is illustrated schematically in Fig. 10.3.

ESI is a soft ionization method wherein molecules are largely spared from fragmentation during the ionization process. Multiply charged ions ($M + nH^{n+}$) are often observed, which is particularly advantageous for the analysis of large macromolecules such as peptides and proteins. The multiple charging reduces the *m/z* value of the molecule to a value that falls within a range that is acceptable to mass analyzers such as quadrupoles and ion traps. In 2002 Fenn [8] was awarded the Nobel Prize in Chemistry for the development of this revolutionary technique. Fenn shared the Nobel Prize with Tanaka who was recognized for the development of laser desorption ionization, which is discussed in a subsequent section of this chapter.

Atmospheric pressure chemical ionization and atmospheric pressure photoionization

APCI is a mode of ionization that is similar to ESI in that ionization takes place at atmospheric pressure, entails nebulization and desolvation, and incorporates an interface design, which is similar to that of ESI. A key difference between APCI and ESI is that ESI is based on ions that are formed in solution, whereas APCI uses ions that are formed in the gas phase. In APCI, instead of high voltage being applied to the inlet capillary, mobile phase from the separation device is evaporated within a heater, and the vapor passes a needle with an applied current, which generates a corona discharge for the formation of stable reaction ions. These ion−molecule reactions involve several patterns such as proton-transfer reactions and electrophilic addition reactions. The high heat (400°C) used with APCI can cause pyrolysis, resulting in

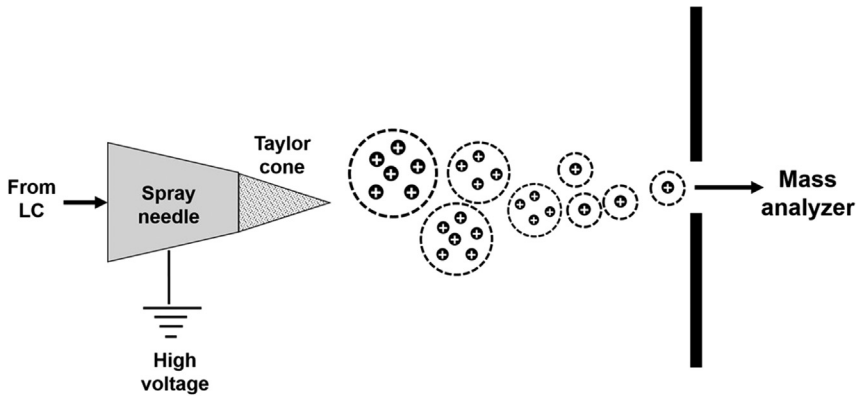


FIGURE 10.3 Electrospray ionization.

the loss of metabolically induced modifications to the analytes of interest. Hence, APCI is best suited for the analysis of small, heat-stable molecules.

APPI is not a mode of ionization that is used frequently in clinical chemistry; however, it is a complementary ionization approach to ESI or APCI. Instead of using a corona discharge needle to generate ions in the gas phase, an ultraviolet (UV) light is used as the source of ionization. To increase the efficiency of analyte ionization, toluene or acetone can be infused coaxially to the nebulizer to provide a source of ions that participate in charge or proton transfer to the analyte molecules. Similar to APCI, ionization occurs after the liquid sample is heated to a gaseous state, which has the benefit of greatly reducing possible matrix effects. Compared with APCI, APPI is more useful for the analysis of compounds with very low polarity.

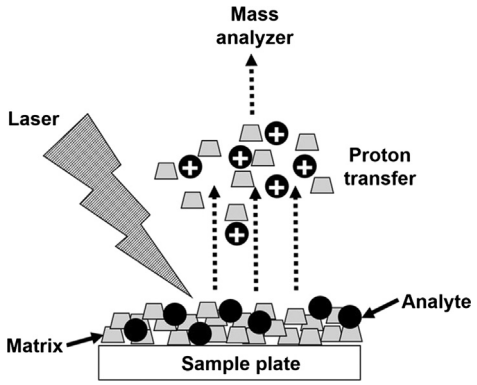


FIGURE 10.4 Matrix-assisted laser desorption ionization. Matrix and analyte are cocrystallized onto a solid support surface. The matrix absorbs at or near the laser wavelength, which causes the matrix molecules to sublime from the surface. Analyte molecules are carried with them, and ionization takes place in the gas phase by proton transfer. The ions are then sampled into the mass analyzer.

Inductively coupled plasma ionization

ICP is an atmospheric pressure ionization method, but unlike the previously mentioned atmospheric pressure ionization methods (ESI, APCI, and APPI), ICP is a hard ionization method, which results in complete sample atomization during sample ionization. An ICP source consists of a sample introduction system (nebulizer and spray chamber that provides the means by which the sample is transmitted into the instrument), an ICP torch, and a radio frequency (RF) coil for the generation of argon plasma, which serves as the ion source, and an interface, which connects the source to the mass spectrometer. In the clinical laboratory, ICP is commonly used for trace element analysis. Among the advantages of ICP are a wide dynamic range and the ability to detect elements at a level of parts per trillion.

Matrix-assisted laser desorption ionization

MALDI is a soft ionization technique that enables the analysis of large, nonvolatile molecules, including proteins, peptides, carbohydrates, and oligonucleotides, from a solid-state phase directly into the gas phase. The first reports of high mass ions $>10{,}000$ m/z that were recorded using MALDI MS were published by two groups in 1988 who reported the detection of molecular ions at m/z 34,529 from carboxypeptidase-A and the detection of molecular ions from bovine serum albumin at m/z 66,750 [9,10]. Tanka [11] was awarded the Nobel Prize in Chemistry in 2002 for his development of laser desorption ionization. As mentioned in a previous section, Tanaka shared this award with Fenn.

The steps entailed in MALDI are as follows: (1) the analyte of interest is suspended or dissolved in a matrix; (2) the analyte cocrystallizes with the matrix as the solvent evaporates; (3) a UV laser (typically a nitrogen laser) irradiates the surface; (4) the matrix absorbs the UV light and ionization occurs; and (5) ions are accelerated and introduced into the mass spectrometer (Fig. 10.4). Several types of matrices are available for use with MALDI. Most matrices are small organic compounds that are able to absorb the energy from UV lasers and gently transfer the energy to the sample, preventing decomposition and promoting the ionization process. Most matrices are mixed with samples at molar ratios of 1000:1 to 10,000:1 matrix:sample.

Laser pulses are targeted at specific areas on the sample plate, containing the matrix cocrystallized sample. Time-of-flight (TOF) mass analyzers are typically used with MALDI ionization sources, because each laser pulse generates a set of ions that can be delivered to the mass analyzer in discrete groups, hence providing distinct start events for the timing of ion flight for each group of ions.

Desorption electrospray ionization and direct analysis in real time

DESI and direct analysis in real time (DART) are ionization methods that enable the generation of ions from surfaces at atmospheric pressure. Before the development of these methods, samples analyzed by MS were not in the free ambient environment nor could they be subjected to processing actions or conditions, while mass spectra were being recorded [12]. In DESI applications, ionization occurs by spraying the sample with an electrically charged aqueous mist. The ions are then released from the surface, and they are transported through air at atmospheric pressure prior to reaching the atmospheric interface of the mass spectrometer. In contrast to the liquid

spray used by DESI, the ionizing gas from the DART ion source contains a dry gas stream with excited state species [13].

Minimal sample preparation, relative ease of operation, and low cost are primary advantages of ambient ionization methods such as DESI and DART. These methods have several clinical applications including dried blood spot testing for therapeutic drug monitoring and lipid profile-based identification of cancerous tissues. Although there are challenges related to sensitivity and matrix effects, novel developments in ambient ionization techniques have created the potential for MS to be included in point-of-care diagnosis.

Emerging ionization methods of interest

Considerable advances in direct sampling methods and ambient analysis have been made in recent years. A few of these advances are highlighted here. Rapid evaporative ionization mass spectrometry (REIMS) allows the direct characterization of biological tissues, food, and environmental samples without the requirement for sample preparation in near real time. In this technique, the aerosol generated from the evaporation of the sample by Joule heating or laser irradiation is introduced into the inlet of a mass spectrometer. REIMS has been demonstrated to have promise in the fields of intrasurgical tissue classification, bacterial identification, and the rapid profiling of cell lines [14].

The intelligent knife (iKnife) was developed by Takats, and it is based on REIMS technology [15]. The iKnife is intended to be used as an alternative to frozen section histology by enhancing near real-time intraoperative decision-making based on surgical margin characterization. By combining a tissue dissection tool (handheld electrosurgical device) with an identification system (MS and a spectral database) via the analysis of the smoke from evaporating tissue during resection, the iKnife has a potential clinical use for patient diagnosis. A clinical trial in the United Kingdom is ongoing to test the iKnife's discriminatory ability and to establish whether the improved accuracy of tissue detection could have a positive impact on progression-free survival and overall survival in breast cancer patients [16].

The MasSpec Pen was developed by researchers at the University of Texas at Austin for use as a clinical and intraoperative device for ex vivo and in vivo cancer diagnosis [17]. Similar to REIMS, it is based on ambient ionization, but it differs from the iKnife method in that it is a nondestructive technique, and biomolecules are extracted from tissues and delivered to the inlet of the mass spectrometer using discrete water droplets as opposed to smoke.

Paper spray MS was developed as a direct sampling ionization method for the MS analysis of complex mixtures, including dried blood spots for therapeutic drug monitoring applications and illicit drug detection in raw urine specimens [18,19]. The method of ionization in paper spray applications is based on electrospray. Ions from the analyte of interest are generated by applying a high voltage to a paper triangle wetted with a small volume ($<10\,\mu L$) of solution. Parameters such as paper thickness, mechanical resistance, surface chemistry, and fiber morphology influence the performance of paper spray, and various functionalized types of paper have been developed for the analysis of specific targets. The potential benefits of paper spray MS in a clinical laboratory include rapid point-of-care analysis.

Coated blade spray (CBS) MS is a solid-phase microextraction-based technology that can be directly coupled to MS to enable the rapid qualitative and quantitative analysis of complex matrices [20]. Combining microsample preparation with ambient ionization, CBS MS uses a polymeric adsorbent particle-coated stainless steel sheet, which functions as a solid-substrate ESI source. Contrary to paper spray MS, analytes of interest are extracted or enriched from a sample in CBS applications.

Laser diode thermal desorption (LDTD)-MS is another direct sampling technique. It was developed primarily to increase analytical throughput by eliminating the need for chromatographic separation prior to MS detection. In this method, a small volume of sample ($\sim 2\,\mu L$) is pipetted into a well of a specially designed 96- or 384-well metallic plate. An infrared laser thermally desorbs the analytes by evaporation of an organic solvent. The desorption process releases neutral gas-phase molecules, which are carried by nitrogen gas through an ion transfer tube into an atmospheric pressure ionization source where they are ionized. In addition to its current applications in the characterization of illicit drugs, explosives, and pesticides, LDTD-MS is being applied to the quantification of trace active pharmaceutical ingredients and their metabolites in blood [21].

Mass analyzers

The operation of mass spectrometers is based on the dynamics of charged particles, or ions, in electric and magnetic fields in a vacuum. These charged particles are separated as a function of their mass and charge in the mass analyzer component of the mass spectrometer. Following their separation, the ions are delivered to the detector.

There are several types of mass analyzers, and they can be broadly classified into two groups: beam-type instruments and trapping-type instruments. In beam-type mass analyzers, ions are delivered from the ion source to

the mass analyzer and then to the detector in a single beam. Comparatively, in trapping-type mass analyzers, ions are retained in a spatially confined region of the mass analyzer for a defined period of time (which can be as short as several milliseconds) prior to being sent to the detector.

The main beam-type mass analyzers are the quadrupole and TOF. Although magnetic sector mass analyzers are also beam-type mass analyzers, they are seldom used in the clinical laboratory and will consequently only be mentioned briefly in this section. The primary trapping-type mass analyzers are the quadrupole ion trap, linear ion trap, ion cyclotron resonance (ICR), and orbitrap.

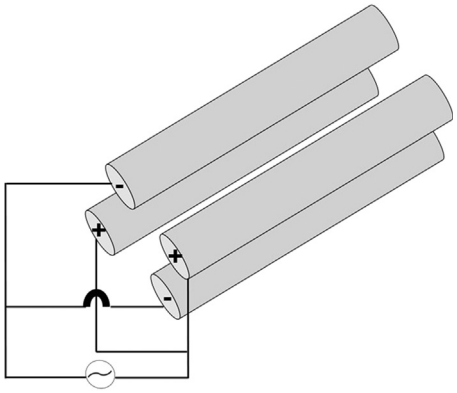


FIGURE 10.5 Quadrupole mass analyzer.

Beam-type design

Magnetic sector

A magnetic sector mass analyzer utilizes an electric or magnetic field to alter the direction of the paths of ions that are accelerated through the mass analyzer. As the ions are deflected toward a detector located at a 90-degree angle from the ion source, the extent of their deflection is based on their *m/z* ratios with lighter ions being deflected to a greater extent compared with heavier ions. Ion separation in magnetic sector analyzers can be described by the following equation:

$$m/z = \frac{B^2 r^2}{2V} \tag{10.1}$$

where B is the strength of the magnetic field, r is the radius, and V is the voltage. Either a narrow range of *m/z* values can be selected by the mass analyzer or a broad range of *m/z* values can be scanned before the ions are transmitted to the detector. The high resolution and mass accuracy of these detectors correlate with their high cost and complex operating principles.

Quadrupole

Quadrupole mass analyzers are sometimes referred to as mass filters due to their ability to select ions of a single *m/z* value for analysis. Although they are inferior to magnetic sector instruments in terms of sensitivity, resolution, and upper mass range, their ease of use, relative affordability, compact size, and ability to be readily interfaced with GC or LC systems have rendered them the most widely used mass spectrometers in the clinical laboratory. As their name suggests, quadrupole mass analyzers comprise four parallel cylindrical metal rods (electrodes with a hyperbolic interior surface) that are positioned equidistant from a center axis. The rods are charged by direct current (DC) and RF voltages, with the opposite pairs of rods carrying like charges (Fig. 10.5). Oscillating electrical fields are used to selectively stabilize or destabilize ions, as they pass through an RF quadrupole field. Each opposing rod pair is connected electrically, and an RF alternating current voltage is applied between the rods. A DC voltage is then superimposed on the RF voltage, which causes the ions to adopt an irregular, oscillatory trajectory, as they travel through the 3-D channel between the rods in a corkscrew path. Only ions with a selected *m/z* value are able to achieve a stable trajectory, which allows them to reach the detector. Ions with other *m/z* values have an unstable trajectory and consequently collide with the rods, or are expelled radially from the analyzer. The selection of ions with a specific *m/z* ratio is achieved by changing the polarity and the voltages that are applied to the rods.

The resolution (R) of quadrupole mass analyzers is directly proportional to the number of RF cycles (n) that an ion undergoes. It therefore follows that increasing the RF frequency is a means by which quadrupole resolution can be optimized. R is defined as follows:

$$R = n^2 = \frac{mf^2L^2}{\mathrm{qV}} \tag{10.2}$$

where m is the mass, f is the RF voltage frequency, L is the rod length, and qV is the kinetic energy (KE) of the ions.

Time of flight

TOF mass analyzers are often used in conjunction with MALDI ionization sources. TOF MS is known as a scanning technique wherein a full mass spectrum is acquired as a snapshot instead of by sequentially stepping through a series of *m/z* values while acquiring the data.

When KE, or force, is imparted to a group of ions with various *m/z* values by the application of an electric field, the ions traverse a path from the ion source to the detector in a flight tube in an amount of time that is

directly related to their *m/z*. Ion velocity (v) is inversely related to the square root of the *m/z*:

$$v = \sqrt{(2V)(z/m)} \tag{10.3}$$

where V represents the voltage. If an equivalent amount of KE, or force, is imparted to all ions, ion velocity will depend only on ion mass:

$$\mathrm{KE} = \frac{mv^2}{2} \tag{10.4}$$

The resolution of early TOF mass analyzers was relatively poor, as the KE for ions with identical *m/z* values was not always equivalent. The development of delayed extraction and the reflectron greatly increased the resolving power of TOF mass analyzers [22]. In the context of MALDI-TOF, delayed extraction involves the inclusion of a short interval between ion creation by the laser and ion acceleration by the application of an electric field, which permits the dense plume of MALDI-generated ions to dissipate before it is accelerated out of the ion source. Consequently, the broadening of ion velocity distribution from collisional processes in the ion source is reduced. In addition, ion arrival time distributions at the detector are narrower, and mass resolution is enhanced. The reflectron compensates for the broadening in the range of flight times of ions with identical *m/z* values by focusing the ion packets in space and time at the detector.

Trapping mass spectrometers

Quadrupole ion trap

Quadrupole ion trap mass analyzers operate based on the same physical principles as quadrupole mass analyzers, but the ions are trapped in stable orbits in a 3-D chamber and are then sequentially ejected. The quadrupole ion trap mass analyzer consists of three hyperbolic electrodes: a donut-shaped ring electrode, an entrance endcap electrode, and the exit endcap electrode (Fig. 10.6—quadrupole ion trap). These electrodes form a cavity in which it is possible to trap (store) and analyze the ions. Both endcap electrodes have a hole in their center through which the ions can travel, and the ring electrode is located midway between the two endcap electrodes.

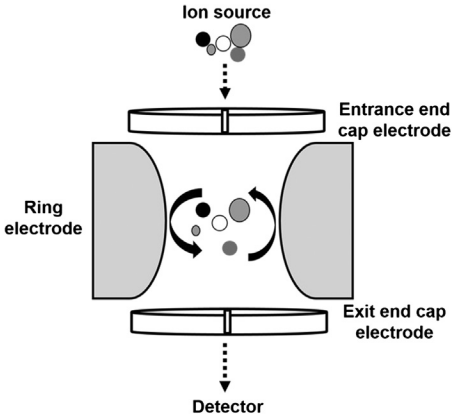


FIGURE 10.6 Quadrupole ion trap mass analyzer.

Ions enter the quadrupole ion trap through the entrance endcap electrode. The ions are then trapped in the space between the electrodes by AC (oscillating, nonstatic) and DC (nonoscillating, static) electric fields. Various voltages are applied to the electrodes to trap and eject ions based on their *m/z*. The ring electrode RF potential produces a 3D quadrupolar potential field within the trapping cavity, which traps ions in a stable oscillating trajectory that is confined within the trapping cell. The exact nature of the trajectory depends on the trapping potential and the *m/z* value of the ions. Then, the electrode system potentials are altered to produce instabilities in the ion trajectories, causing the ions to be axially ejected in order of increasing *m/z* value and focused by the exit lens prior to detection. An important distinction between quadrupole mass analyzers and quadrupole ion trap mass analyzers is that quadrupole mass analyzers separate and detect masses by allowing oscillating ions to pass through the quadrupole to reach the detector (beam-type mass analyzer), whereas quadrupole ion trap mass analyzers separate and detect masses by discharging ions with unstable oscillations from the system (trapping-type mass analyzer).

Quadrupole ion trap mass analyzers are versatile instruments that are routinely used for exploratory studies, structural characterization, and qualitative identification. Their advantages include the ability to conduct tandem MS analysis (see the "Tandem mass spectrometers" section) and high sensitivity. However, quadrupole ion traps are not well suited for quantitative analysis due to the possibility of space-charge (ion−ion repulsion) effects resulting from the limited number of charged species that can fit within the trap, which reduces the analyzer's dynamic range.

Linear ion trap

Linear, or 2D, ion trap mass analyzers use a set of quadrupole rods to radially confine ions and a static electric potential on the end electrodes to axially confine ions. Similar to quadrupole mass analyzers, linear ion traps can also be used as selective mass filters. Among the advantages of linear ion traps are increased ion storage capacity, rapid scan times, and versatile modes of operation. Compared with quadrupole ion trap mass analyzers, linear ion traps have a much greater trapping volume, which enables more ions to be trapped before space-charge effects become a matter of concern.

Ion cyclotron resonance

ICR-MS, or FTICR-MS, is based on the principle that ions in a magnetic field follow a circular motion (cyclotroning).

These mass analyzers measure *m/z* values by detecting the image current produced by ions that are accelerated outward from a common center along a spiral path in a magnetic field created by a high-field superconducting magnet (3−12 T). Ions are injected into a Penning trap (static electric/magnetic ion trap), where they form part of a circuit. Detectors at fixed positions measure the electrical signals from the ions passing by them in close proximity, producing cyclical signals that can be deconvoluted by performing a Fourier transform on the signal.

Among the main advantages of ICR-MS are its high sensitivity, high mass accuracy, and superior resolution (can exceed 1 million) [23]. As ICR mass analyzers are trapping-type analyzers, they are amenable to various modes of MS analysis, including MS^n. However, it should be noted that ion fragmentation is often not required for analyte detection due to the high mass accuracy of the mass analyzer. The disadvantages of ICR mass analyzers include high cost of instrument acquisition and operation, considerable space requirements due to the instrument's large footprint, requirement for a high-field superconducting magnet, relatively long signal acquisition time, safety concerns related to the high magnetic fields, and requirement of highly skilled operators. It is for these reasons that ICR mass analyzers are not amenable for use in routine clinical laboratory applications.

Orbitrap

The orbitrap mass analyzer was commercialized in 2006 [24,25], and it is similar to the ICR mass analyzer, but it does not utilize a magnetic field for trapping. It can readily achieve resolutions up to 100,000 and up to subparts per million mass accuracy, and its dynamic range is up to four orders of magnitude.

The operational principle of this trapping-type mass analyzer is trapping within electrostatic fields. The analyzer itself is a spindle-like central electrode surrounded by a barrel-like outer electrode. The central electrode confines ions, so that they orbit around the electrode and oscillate back and forth along the long axis of the central electrode. This oscillation generates an image current, and the frequency of which depends on the *m/z* value of the individual ions. Mass spectra are obtained in a manner similar to that of FTICR mass analyzers.

Orbitrap mass analyzers are commonly used in proteomic biomarker discovery applications in research laboratory settings [26], and they have great potential for use in clinical toxicology applications based on their high resolution and mass accuracy, which are beneficial to definitive analyte identification [27]. Orbitrap mass analyzers can be readily interfaced with LC systems, thus enabling the incorporation of the use of retention time along with accurate mass to aid compound identification.

Ion mobility

Although ion mobility mass spectrometers (IMSs) are not technically mass analyzers, they are worth mentioning in this section. IMS requires analyte ionization prior to detection, but instead of separating ions based on their *m/z*, ions are separated based on their mobility in an electric field. An IMS can operate at atmospheric pressure or reduced pressure, and in mass spectrometer configurations, an IMS can be situated upstream of the first mass analyzer or after one stage of mass analysis. In the clinical laboratory, IMS has been used for lipoprotein subfraction categorization [28].

Tandem mass spectrometers

In addition to being configured with only one mass analyzer, mass spectrometers can be configured with two or more mass analyzers arranged sequentially, with a collision cell located in between them to enable tandem MS, or MS/MS, analysis. Tandem MS has great utility for several applications including compound identification, the characterization of complex structures, and quantitative analysis.

MS/MS can be achieved using many different analyzer configurations. For beam-type mass analyzers, MS/MS analysis is accomplished via "tandem in space" wherein the analyzers are arranged in a linear sequence. This is in comparison with trapping-type mass analyzers wherein MS/MS analysis is accomplished via "tandem in time"—ions are retained in one region of space, while the precursor ion is selected and dissociated and the fragment ion(s) are analyzed sequentially in time.

In the most widely used configuration of trapping mass analyzers for MS/MS analysis, the first mass analyzer is used to select a precursor ion of a specific *m/z* value, which is an unfragmented ion. The precursor ion then passes into a collision cell where it collides with inert background gas molecules (often argon, helium, or nitrogen) to induce fragmentation, resulting in the formation of product or fragment ions. Mass analysis of the fragment ions is conducted in the second mass analyzer. When coupled with chromatographic separation, MS/MS applications have great selectivity owing to the ability to characterize compounds based on their chromatographic retention time, precursor ion mass, and product ion mass.

Although collision-induced dissociation (CID) is perhaps the most commonly used method to fragment gas-phase ions in MS, several other methods of fragmentation exist including electron transfer dissociation (ETD), electron capture dissociation (ECD), infrared multiphoton dissociation (IRMPD), and higher-energy C-trap (or collisional) dissociation (HCD). The fundamentally similar aspect of these methods of fragmentation is the dissociation of energetically unstable molecular ions, resulting

in unique mass spectral patterns that can often be compared with the spectra in well-curated spectral libraries for the identification of the analyte(s) of interest.

CID refers to the process whereby an ion and a neutral gas (helium, nitrogen, or argon) collide, resulting in fragmentation. CID is useful in applications requiring structural determination. HCD is a variation of CID specific to orbitrap mass spectrometers wherein a higher RF voltage is used to retain ions in a C-trap. The HCD cell is used to fragment the ions, after which they are accelerated and stored inside of the C-trap. The ions are then injected into and separated inside of the orbitrap mass analyzer based on differences in their rotational frequency. HCD is not limited by the low mass cutoff of CID fragmentation and is therefore advantageous for applications entailing isobaric tag-based quantification.

ETD is primarily used for sequence analysis of peptides and proteins including the identification of labile posttranslational modifications. This method of fragmentation requires multiply charged ($z>2$) gas-phase cations, and it is commonly limited to use with an ESI source. After the cations are captured in an ion trap, they react with an anion such as the polycyclic aromatic hydrocarbon molecule, fluoranthene. The capture of an electron from the anion yields an unstable cation radical. The cation radial then breaks into two fragments, an N-terminal c-ion and a C-terminal z-ion. This cleavage occurs randomly between the N−Cα bonds of any two amino acids except proline on the peptide backbone. Similar to ETD, ECD induces fragmentation of higher charge state cationic molecules by transforming their electric potential energy to KE. Although ECD is used primarily with FTICR MS, it also has applications for quadrupole ion trap mass spectrometers.

IRMPD is another fragmentation method that is most often used with FTICR MS. It entails the introduction of a laser beam into an ICR cell. IRMPD fragmentation is based on the absorption by a given ion of multiple infrared photons from a tunable laser. Precursor ions become excited into more energetic vibrational states until a bond is broken, resulting in fragment ion spectra that can contain a wealth of molecular structure information.

Many different scan functions are possible with MS/MS, and they are summarized in Table 10.1. Product ion scanning entails the selection of a precursor ion of a given *m/z* in the first mass analyzer (termed MS1 or Q1) followed by a full scan of the complete spectrum of fragment or target ions. This mode of analysis is frequently used for peptide sequencing analysis. Precursor ion scanning is essentially the reverse of product ion scanning. In this mode of analysis, the second mass analyzer selects a product ion of a specific *m/z*, and the first mass analyzer scans the spectrum of precursor ions to enable the determination of the precursor ions that produce a specific product ion. In the neutral loss mode, both mass analyzers are scanned simultaneously for a constant difference between the transmitted *m/z* values. A signal is recorded only if an ion undergoes fragmentation, producing a neutral ion that is equivalent to the mass difference of interest. The most commonly used scan function in MS/MS applications is multiple (or selected) reaction monitoring (MRM), and it is performed using a triple quadrupole mass spectrometer in which two sets of quadrupoles are separated by an RF-only quadrupole, which functions as a collision cell and passes all ions regardless of *m/z* to the last quadrupole. In MRM acquisition, an ion corresponding to the intact molecular ion of the analyte of interest is selected in the first mass analyzer followed by

TABLE 10.1 Various scan modes entailing MS/MS fragmentation.

Scan mode	Description
Product ion scanning	• Q1 analyzer set to allow only the transmission of 1 *m/z* • Precursor ion collided with inert gas to form product ions • Product ions scanned in Q3 analyzer
Precursor ion scanning	• Q1 analyzer scanned • Precursor ion collided with inert gas to form product ions • Q3 analyzer set to allow only a single fragment ion of 1 *m/z* to pass
Neutral loss	• Q1 and Q3 scanned • Precursor ions collide with inert gas to form product ions • Q3 mass analyzer offset by neutral loss mass of interest • Only ions yielding product ions with the loss of interest are detected
MRM	• Q1 analyzer set to allow the transmission of precursor ion of interest • Fragmentation in collision cell produces product ions • Q3 analyzer set to allow transmission of product ion(s) of interest

MRM, Multiple reaction monitoring.

fragmentation in the collision cell to produce a range of product ions. One or more of these product ions can then be selected for quantitative analysis. MRM acquisition methods are particularly sensitive and specific.

Tandem mass spectrometer configurations are not limited to identical mass analyzers. Hybrid instruments exist, which combine two or more different types of analyzers. Among the most common configurations of hybrid instruments are quadrupole and TOF analyzers, as well as quadrupole and linear ion trap analyzers.

Detectors and processors

In mass spectrometers, after ions travel from the ion source to the mass analyzer(s), they reach the detector where a signal is emitted in response. Ions are converted to a current, and the current is amplified to increase the gain, which is a measure of the intensity of the ions that strike the detector.

Electron multipliers are the most commonly used detectors for mass spectrometers, with the exception of ICR, orbitrap, and some ICP mass spectrometers. The predominant types of electron multipliers are discrete dynode multipliers, continuous dynode electron multipliers, and microchannel plate electron multipliers. In discrete dynode electron multipliers, a series of dynodes are used to amplify the initial signal, thereby generating an electron cascade. After an ion strikes the first dynode, at least one electron is ejected from the dynode surface. This process is repeated for the remaining dynodes, resulting in a gain of 10^4-10^8 electrons.

Detection in ICR and orbitrap mass spectrometers is based on image current detection. Instead of ions being detected by striking a detector such as an electron multiplier, ions induce sinusoidal image current on detection electrodes, as they undergo cyclotron motion. Ions with different *m/z* values oscillate with different cyclotron frequencies. The signal collected by the electrodes is a sum of the various sine waves and is commonly referred to as a transient. Conversion of the transient in the time domain to the frequency domain occurs via Fourier transformation and results in the creation of a mass spectrum.

Following detection of the ions, the raw signal is processed either by the counting of individual pulses that correspond to individual ions, or the signal is converted to a digital representation. Regardless of the downstream data analysis method, when working with raw MS data, it is critical to differentiate between true signal, baseline, and noise. When tuning a mass spectrometer based only on the increase of the signal of a selected number of *m/z* values, an increase in noise often occurs concomitantly with an increase in signal. The relationship between increased noise and signal is not 1:1. Thus it is important to determine the signal-to-noise ratio, which is of fundamental importance when evaluating limits of detection and quantification.

Clinical mass spectrometry applications

As mentioned at the outset of this chapter, MS is a powerful analytical tool that is providing clinical laboratories with enhanced capabilities. This section will briefly highlight some of the ways in which MS-based applications are being used in the clinical laboratory, but the reader is encouraged to refer to Chapter 21, Applications of mass spectrometry in the clinical laboratory, for a more thorough discussion of this topic.

GC-MS based methods are routinely used by the US National Institute of Standards and Technology (NIST) for the development of definitive methods to quantify standard reference materials and assign accurate concentrations to reference materials. GC-MS is also used in clinical and forensic drug testing and screening for certain inborn errors of metabolism. LC-MS and LC-MS/MS assays are used for the measurement of several clinically relevant analytes including immunosuppressant drugs [29], vitamins [30], antiretroviral drugs [31], thyroid hormones [32], thyroglobulin [33], and steroids [34]. The majority of these assays incorporate the use of MRM due to its specificity, selectivity, and analytical sensitivity.

In the clinical microbiology laboratory, MALDI-TOF MS is now routinely used for rapid microbial identification. The implementation of MALDI-TOF MS in this setting has been shown to have cost savings of >50% compared with standard culture techniques [35]. MALDI-TOF MS is also routinely used in genomics applications for the identification of nucleic acid sequence changes (e.g., mutations, single-nucleotide polymorphisms, insertions/deletions, and alternative splicing), quantitative genomic changes (e.g., copy number variation, gene expression, and allele expression), and identification of modifications (e.g., genomic DNA methylation and post-transcriptional modification of tRNAs and rRNAs) [36,37].

Clinical applications of DESI MS-based methods include the surgical iKnife using REIMS [38] and the MasSpec Pen [17]. These devices were described earlier in the "Emerging ionization methods of interest" section.

Challenges and opportunities in the implementation of clinical mass spectrometry methods

The challenges in implementing clinical MS methods are well-known and include the lack of automation, manual nature of the assays, relatively low sample throughput, high capital expense ($200,000−$500,000 plus associated

construction/renovation costs), high level of required technical proficiency, limited assay standardization, and availability of commercial material/traceability [39]. Careful attention to the following parameters when developing MS assays will mitigate some of the drawbacks that lead to poor assay performance: test selection, matrix selection, analyte stability, establishment of reference intervals, and instrument downtimes.

The lack of optimization of some of the following essential features of MS can significantly compromise data quality and hence patient safety: sample preparation, prefractionation, ionization, ion manipulation, signal readout, and data processing. Accordingly, systemic quality management and sustained vigilance are required for MS-based methods that are applied in the clinical laboratory. The fundamental aspects of quality in MS methods applied in the clinical laboratory can be summarized as follows: (1) method design and optimization; (2) appropriate handling of the metadata associated with each run and each peak as postimplementation surveillance; (3) careful management of all steps of the analytical process to avoid gross handling errors; (4) adequate training of technicians including competency verification according to widely accepted quality management standards; (5) a proactive maintenance plan to limit the within- and between-instrument variations; and (6) continued quality verification based on the analysis of control samples and proficiency testing of samples with reliable assigned target concentrations [40].

The Clinical and Laboratory Standards Institute (CLSI) C50A [41] guidance document "Mass Spectrometry in the Clinical Laboratory: General Principles and Guidance" provides a general overview of MS and the principles that dictate its application in the clinical laboratory. The CLSI C62A [42] guidance document "Liquid Chromatography-Mass Spectrometry Methods" provides guidance to the clinical laboratorian for the reduction of interlaboratory variance and the evaluation of interferences, assay performance, and other relevant characteristics of clinical assays. Clinical laboratories performing LC-MS-based tests should be aware of the five directly applicable Clinical Laboratory Improvement Amendments requirements and the associated guidelines: facility administration, quality systems, proficiency testing, personnel, and inspection [43].

Indeed, MS is a highly sensitive and selective technique that can be used for the quantitative analysis of several clinically relevant analytes. Some of these highly specific assays can overcome the limitations of immunoassays, including cross-reactivity, low analytical specificity, and limited dynamic range [44]. Additional advantages of clinical MS methods include rapid assay development compared with immunoassays, potential cost savings on high volume send-out tests, the ability to multiplex multiple analytes in a single method, and the ability to analyze a wide variety of analytes including drugs, hormones, small molecules, carbohydrates, nucleic acids, and proteins. For many clinical applications, there is a clear benefit to the implementation of MS; however, there is considerable room for improvement of the application of this powerful technology in the clinical laboratory. Many of these desirable areas of improvement are related to enhanced automation, more FDA-approved MS-based platforms, and the commercial availability of NIST-traceable assay kits, calibrators, and QC materials. MS continues to be an ever-evolving technology due to the rapid pace of advances in instrumentation. Realization of the full potential of MS in the clinical laboratory requires a thorough understanding of not only the vast opportunities but also the limitations of the technology.

Acknowledgments

The author gratefully acknowledges the original contributions to this chapter in the third edition of *Contemporary Practice in Clinical Chemistry* by Autumn Breaud, Joely A. Straseski, and William Clarke, on which portions of this chapter are based.

References

[1] J.M. Thomas, J. J. Thomson: winner of the Nobel Prize for Physics 1906, Angew. Chem. Int. Ed. Engl. 45 (41) (2006) 6797–6800.

[2] F.W. McLafferty, A century of progress in molecular mass spectrometry, Annu. Rev. Anal. Chem. (Palo Alto Calif.) 4 (2011) 1–22.

[3] P.J. Jannetto, R.L. Fitzgerald, Effective use of mass spectrometry in the clinical laboratory, Clin. Chem. 62 (1) (2016) 92–98.

[4] F.W. McLafferty, F. Tureek, Interpretation of Mass Spectra, fourth ed., University Science Books, Herndon, 1993, pp. 371.

[5] T.M. Annesley, Ion suppression in mass spectrometry, Clin. Chem. 49 (7) (2003) 1041–1044.

[6] B.K. Matuszewski, M.L. Constanzer, C.M. Chavez-Eng, Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC-MS/MS, Anal. Chem. 75 (13) (2003) 3019–3030.

[7] H.H. Maurer, Current role of liquid chromatography-mass spectrometry in clinical and forensic toxicology, Anal. Bioanal. Chem. 388 (7) (2007) 1315–1325.

[8] J.B. Fenn, et al., Electrospray ionization for mass spectrometry of large biomolecules, Science 246 (4926) (1989) 64–71.

[9] M. Karas, F. Hillenkamp, Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons, Anal. Chem. 60 (20) (1988) 2299–2301.

[10] K. Tanaka, et al., Protein and polymer analyses up to *m/z* 100 000 by laser ionization time-of-flight mass spectrometry, Rapid Commun. Mass. Spectrom. 2 (1988) 151–153.

[11] K. Tanaka, The origin of macromolecule ionization by laser irradiation (Nobel lecture), Angew. Chem. Int. Ed. Engl. 42 (33) (2003) 3860–3870.

[12] Z. Takats, J.M. Wiseman, R.G. Cooks, Ambient mass spectrometry using desorption electrospray ionization (DESI): instrumentation, mechanisms and applications in forensics, chemistry, and biology, J. Mass. Spectrom. 40 (10) (2005) 1261–1275.
[13] R.B. Cody, J.A. Laramee, H.D. Durst, Versatile new ion source for the analysis of materials in open air under ambient conditions, Anal. Chem. 77 (8) (2005) 2297–2302.
[14] E.A. Jones, et al., Matrix assisted rapid evaporative ionization mass spectrometry, Anal. Chem. 91 (15) (2019) 9784–9791.
[15] J. Balog, et al., Intraoperative tissue identification using rapid evaporative ionization mass spectrometry, Sci. Transl. Med. 5 (194) (2013) 194ra93.
[16] D. Leff, Z. Takats, Real time tissue characterization using mass spectrometry REI-EXCISE iKnife study (REI-EXCISE), in: U.S. National Library of Medicine Clinical Trials.gov, Imperial College London, 2018.
[17] J. Zhang, et al., Nondestructive tissue analysis for ex vivo and in vivo cancer diagnosis using a handheld mass spectrometry system, Sci. Transl. Med. 9 (406) (2017).
[18] N.E. Manicke, et al., Quantitative analysis of therapeutic drugs in dried blood spot samples by paper spray mass spectrometry: an avenue to therapeutic drug monitoring, J. Am. Soc. Mass. Spectrom. 22 (9) (2011) 1501–1507.
[19] H. Wang, et al., Paper spray for direct analysis of complex mixtures using mass spectrometry, Angew. Chem. Int. Ed. Engl. 49 (5) (2010) 877–880.
[20] G.A. Gómez-Ríos, M. Tascon, J. Pawliszyn, Coated blade spray: shifting the paradigm of direct sample introduction to MS, Bioanalysis 10 (4) (2018) 257–271.
[21] C.H. Wang, et al., Solid phase microextraction combined with thermal-desorption electrospray ionization mass spectrometry for high-throughput pharmacokinetics assays, Anal. Chim. Acta 1021 (2018) 60–68.
[22] T.J. Cornish, R.J. Cotter, A curved field reflectron time-of-flight mass spectrometer for the simultaneous focusing of metastable product ions, Rapid Commun. Mass. Spectrom. 8 (9) (1994) 781–785.
[23] A.G. Marshall, C.L. Hendrickson, G.S. Jackson, Fourier transform ion cyclotron resonance mass spectrometry: a primer, Mass. Spectrom. Rev. 17 (1) (1998) 1–35.
[24] A. Makarov, et al., Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer, J. Am. Soc. Mass. Spectrom. 17 (7) (2006) 977–982.
[25] R.H. Perry, R.G. Cooks, R.J. Noll, Orbitrap mass spectrometry: instrumentation, ion motion and applications, Mass. Spectrom. Rev. 27 (6) (2008) 661–699.
[26] R. Aebersold, et al., How many human proteoforms are there? Nat. Chem. Biol. 14 (3) (2018) 206–214.
[27] A.H. Wu, et al., Role of liquid chromatography-high-resolution mass spectrometry (LC-HR/MS) in clinical toxicology, Clin. Toxicol. (Phila.) 50 (8) (2012) 733–742.
[28] M.P. Caulfield, et al., Direct determination of lipoprotein particle sizes and concentrations by ion mobility analysis, Clin. Chem. 54 (8) (2008) 1307–1316.
[29] R.W. Peake, et al., Multicenter evaluation of the thermo scientific prelude for measurement of immunosuppressant drugs using sample preparation liquid chromatography-tandem mass spectrometry, Ther. Drug. Monit. 37 (2) (2015) 161–171.
[30] N. Heureux, Vitamin D testing-where are we and what is on the horizon? Adv. Clin. Chem. 78 (2017) 59–101.
[31] T. Koal, et al., Quantification of antiretroviral drugs in dried blood spot samples by means of liquid chromatography/tandem mass spectrometry, Rapid Commun. Mass. Spectrom. 19 (21) (2005) 2995–3001.
[32] B. Yue, et al., Free thyroid hormones in serum by direct equilibrium dialysis and online solid-phase extraction--liquid chromatography/tandem mass spectrometry, Clin. Chem. 54 (4) (2008) 642–651.
[33] A.N. Hoofnagle, et al., Quantification of thyroglobulin, a low-abundance serum protein, by immunoaffinity peptide enrichment and tandem mass spectrometry, Clin. Chem. 54 (11) (2008) 1796–1804.
[34] S.J. Soldin, O.P. Soldin, Steroid hormone analysis by tandem mass spectrometry, Clin. Chem. 55 (6) (2009) 1061–1066.
[35] K.E. Tan, et al., Prospective evaluation of a matrix-assisted laser desorption ionization-time of flight mass spectrometry system in a hospital clinical microbiology laboratory for identification of bacteria and yeasts: a bench-by-bench study for assessing the impact on time to identification and cost-effectiveness, J. Clin. Microbiol. 50 (10) (2012) 3301–3308.
[36] X. Gao, et al., MALDI mass spectrometry for nucleic acid analysis, Top. Curr. Chem. 331 (2013) 55–77.
[37] K. Meyer, P.M. Ueland, Use of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for multiplex genotyping, Adv. Clin. Chem. 53 (2011) 1–29.
[38] D.L. Phelps, et al., The surgical intelligent knife distinguishes normal, borderline and malignant gynaecological tissues using rapid evaporative ionisation mass spectrometry (REIMS), Br. J. Cancer 118 (10) (2018) 1349–1358.
[39] W. Clarke, J.M. Rhea, R. Molinaro, Challenges in implementing clinical liquid chromatography-tandem mass spectrometry methods--the light at the end of the tunnel, J. Mass. Spectrom. 48 (7) (2013) 755–767.
[40] M. Vogeser, C. Seger, Quality management in clinical application of mass spectrometry measurement systems, Clin. Biochem. 49 (13–14) (2016) 947–954.
[41] Clinical and Laboratory Standards Institute, Liquid Chromatography-Mass Spectrometry Methods; Approved Guideline. C62-A, CLSI, Wayne, PA, 2014.
[42] Clinical and Laboratory Standards Institute, Mass Spectrometry in the Clinical Laboratory: General Principles and Guidance; Approved Guideline. C50-A, CLSI, Wayne, PA, 2007.
[43] K.L. Lynch, Accreditation and quality assurance for clinical liquid chromatography-mass spectrometry laboratories, Clin. Lab. Med. 38 (3) (2018) 515–526.
[44] A.N. Hoofnagle, M.H. Wener, The fundamental flaws of immunoassays and potential solutions using tandem mass spectrometry, J. Immunol. Methods 347 (1–2) (2009) 3–11.

Further reading

R.J. Cotter, Time-of-Flight Mass Spectrometry: Instrumentation and Applications in Biological Research, American Chemical Society, Washington, DC, 1997, pp. 326.

C. Dass, Fundamentals of Contemporary Mass Spectrometry, John Wiley & Sons, Inc, Hoboken, 2007, pp. 585.

P.H. Dawson, Quadrupole Mass Spectrometry and Its Applications, Elsevier, New York, 1976, pp. 372.

F.W. McLafferty, F. Tureek, Interpretation of Mass Spectra, fourth ed., University Science Books, Herndon, 1993, pp. 371.

M.C. McMaster, LC/MS: A Practical User's Guide, John Wiley & Sons, Inc, Hoboken, 2005, pp. 165.

A.L. Rockwood, M.M. Kushnir, N.J. Clarke, Mass spectrometry, in: N. Rifai, A.R. Horvath, C.T. Wittwer (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, sixth ed., Elsevier, St. Louis, MO, 2018, pp. 295–323.

G. Siuzdak, Mass analyzers and ion detectors, in: G. Siuzdak (Ed.), Mass Spectrometry for Biotechnology, Academic Press, San Diego, CA, 1996, pp. 32–55.

Self-assessment questions

1. Which of the following is the correct order of mass spectrometry analysis?
   a. MALDI → GC → TOF → processor
   b. LC → ESI → quadrupole ion trap → electron multiplier → processor
   c. ESI → MALDI → MS/MS → detector
   d. LC → DESI → ICR → APPI
   e. GC → TOF/TOF → APPI → processor → electron multiplier
2. Which of the following is not a type of mass analyzer?
   a. MALDI
   b. quadrupole ion trap
   c. linear ion trap
   d. ion cyclotron resonance
   e. TOF
3. Which of the following pairs is incorrectly matched?
   a. electron ionization: hard ionization
   b. quadrupole: corkscrew ion path
   c. chemical ionization: soft ionization
   d. MALDI: soft ionization
   e. APCI: pulsed laser beam
4. Beam type mass analyzers include which of the following?
   a. ICR
   b. TOF
   c. quadrupole ion trap
   d. APCI
   e. linear ion trap
5. What is the principle that governs time-of-flight analyzers?
   a. Heavy ions do not move (fly) as fast as light ions.
   b. Ion flight is measured one ion at a time.
   c. Ion flight time is independent of the ion *m/z*.
   d. Switching voltages allows only selected ions to reach the detector.
   e. Ions following an unstable trajectory will be eliminated.
6. Ion fragmentation can occur within which region(s) of a mass spectrometer?
   a. ion source
   b. vacuum
   c. detector
   d. collision cell
   e. b and c
   f. a and d
7. Which of the following is/are not an advantage of mass spectrometry analysis?
   a. increased selectivity
   b. fully automated instrument systems
   c. few interferences
   d. lack of matrix effects
   e. all of the above
   f. b and d
   g. a and c
8. Which of the following aspects of mass spectrometry is correct?
   a. Resolution can be determined from peak height.
   b. Mass accuracy is defined by the mass difference divided by the true mass.
   c. A mass spectrum is a plot of ion mass versus charge.
   d. Mass-to-charge ratio is defined as $m/\Delta m$.
   e. Ion formation occurs in the mass analyzer.
9. ESI is characterized by which of the following correctly ordered steps?
   a. Taylor cone formation → voltage applied → ions enter MS inlet → desolvation/evaporation
   b. Voltage applied → ions enter MS inlet → desolvation/evaporation → Taylor cone formation
   c. Voltage applied → Taylor cone formation → ions enter MS inlet → desolvation/evaporation
   d. Ions enter MS inlet → voltage applied → desolvation/evaporation → Taylor cone formation
   e. Voltage applied → Taylor cone formation → desolvation/evaporation → ions enter MS inlet
10. Which of the following statements is incorrect?
    a. Tandem mass spectrometry is highly selective because of the analysis of precursor and product ions.
    b. Quadrupole resolution is affected by applied voltages.
    c. ICR mass analyzers are commonly used with MALDI ionization sources.
    d. Only specimens capable of being volatilized to the gas phase are suitable for mass spectrometry analysis.
    e. MS^n is described as multiple, sequential rounds of ion fragmentation and analysis.

Answers

1. b
2. a
3. e
4. b
5. a
6. f
7. f
8. b
9. e
10. c

Chapter 11

Nuclear magnetic resonance technology and clinical applications

Deanna D.H. Franke[1] and Margery A. Connelly[2]

[1]*Technical Specialties, Atrium Health, Corporate Lab Services, NC, United States,* [2]*Laboratory Corporation of America Holdings (LabCorp), NC, United States*

Learning objectives

After completing this chapter, the reader will be able to:

1. Review theory and principles of NMR.
2. Summarize current applications of NMR technology for clinical analysis.
3. List three NMR-based tests that are currently available for reporting of patient results in the clinical laboratory.

Introduction

Nuclear magnetic resonance (NMR) has become a vital tool and transformative technology capable of producing molecular fingerprints for diagnostic use. This chapter will provide an overview of the theory, principles, and clinical applications of NMR spectroscopy. From bench to bedside, advancements in the NMR field have transformed how scientists, and clinical and laboratory professionals interrogate the human body—both physically and biochemically. The past couple of years have seen an expansion of use of NMR in the clinical laboratory, largely due to the high-throughput nature of NMR analyzers that have been designed specifically for this purpose. NMR provides the ability to quantify multiple analytes simultaneously by the acquisition of a single NMR spectrum. Furthermore, the development of powerful new software algorithms facilitates the quantitation of individual analytes while taking into account overlapping signals from the multitude of molecules in complex samples such as plasma and serum.

Theory of nuclear magnetic resonance

"NMR" is a property of the nucleus of an atom. Each positively charged nucleus spins on its own axis and creates a small nuclear magnetic field or magnetic moment (Fig. 11.1A). In the absence of an external magnetic force, nuclear magnetic fields are randomly oriented (Fig. 11.1B). However, when placed in a strong external magnetic field, the magnetic field of each nucleus aligns with the external magnetic field in either the same or the opposite direction, similar to a miniature bar magnet (Fig. 11.1C). When electromagnetic radiation in the radiofrequency range is applied to the sample, the nuclei that were originally aligned in the opposite direction to the external magnetic field flip, and the nuclei are then said to be "in resonance" with the strength of the external magnetic field (Fig. 11.1D). Different isotopes can have a variety of values for nuclear spin angular momentum or quantum number (I): no overall spin or zero spin (for nuclides with even numbers of protons and/or neutrons) or one spin (for nuclides with odd numbers of protons and/or neutrons). The most useful nuclei for NMR spectroscopy are nuclides with paired neutrons plus a proton of spin ½, where I = ½. Nuclei with ½ spin include hydrogen (^{1}H), carbon (^{13}C), fluorine (^{19}F), nitrogen (^{15}N), and phosphorus (^{31}P). These nuclei are "NMR active," as they produce the signals that comprise the NMR spectra and allow for the predictable determination of protons within their functional groups or neighboring substituents. These ½ spin, "NMR active" nuclei are the most common elements in biological specimens and, as such, can be readily analyzed by NMR spectroscopy. Because hydrogen (^{1}H) or proton NMR spectroscopy is the most widely used NMR technique in the clinical laboratory, the focus of this chapter will be on proton NMR spectroscopy.

Proton nuclear magnetic resonance spectroscopy

The underlying principle of proton (^{1}H) NMR spectroscopy is similar to that of magnetic resonance imaging

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00011-9

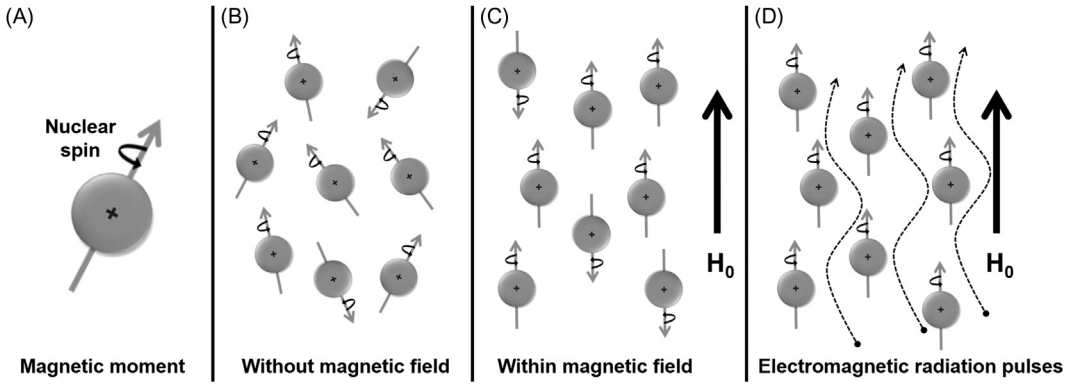


FIGURE 11.1 Nuclear magnetic resonance theory. (A) Nuclei with an unpaired proton have spin and a net magnetic moment. (B) When not in a magnetic field, the nuclear spins are randomly oriented. (C) In an external magnetic field, the nuclei align with the external magnetic field in either the same or the opposite direction. (D) The nuclei become aligned in the same direction and are said to be in resonance with the external magnetic field when subjected to pulses of electromagnetic radiation.

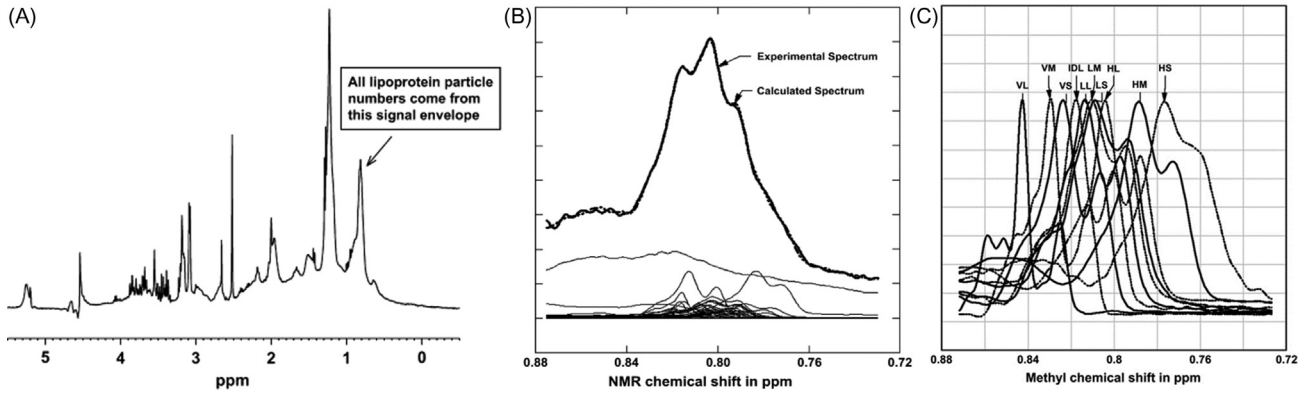


FIGURE 11.2 Proton nuclear magnetic resonance spectrum of plasma. (A) A typical proton nuclear magnetic resonance spectrum of human plasma recorded on a 400-MHz nuclear magnetic resonance clinical analyzer. The signal envelope centered at 0.8 ppm arises from the methyl proton signals of the lipoproteins. (B) A representative deconvolution of the methyl signal (0.7−0.9 ppm). The solid line represents the experimental spectrum, and the dotted line is the calculated sum of the individual lipoprotein subclass signals shown at the bottom. The broad peak in the middle accounts for the nonlipoprotein plasma protein signal. (C) Spectra of several lipoprotein subclasses. The lipoprotein subclasses were isolated by ultracentrifugation and agarose column chromatography and the spectrum for each subclass was acquired in order to illustrate their distinct nuclear magnetic resonance signals. *HL*, large HDL; *HM*, medium HDL; *HS*, small HDL; *IDL, intermediate-density lipoprotein*; *LL*, large LDL; *LM*, medium LDL; *LS*, small LDL; *VL*, Large VLDL; *VM*, medium VLDL; *VS*, small VLDL.

(MRI) [1]. Both involve placing an object or a specimen in a strong external magnetic field, irradiating the object or specimen with harmless radiofrequency pulses and recording the response of the nuclei. While MRI provides anatomic information, NMR spectroscopy generates an electromagnetic spectrum, providing molecular information (Fig. 11.2A). More specifically, ^{1}H NMR spectroscopy exploits the magnetic properties of protons in order to obtain information about both the identity and quantity of hydrogen containing molecules within a specimen. A description of the theoretical basis of ^{1}H NMR and a simplified depiction of an NMR spectrometer can be found in a recent review by Rankin et al. [2]. In brief, the test sample is placed in a strong external magnetic field and electromagnetic radiation, in the form of radiofrequency pulses, is used to excite the protons on the various molecules within the sample. As the protons relax back to equilibrium, the released energy is recorded as an oscillating electromagnetic signal, called the free induction decay (FID). Fourier transformation, a mathematical transformation of the FID signal from time domain to frequency domain, results in a spectrum with intensity plotted on the *y*-axis and chemical shift (δ), in parts per million (ppm), plotted on the *x*-axis (Fig. 11.2B, Fig. 11.2C).

The position in the NMR spectrum, also referred to as the chemical shift, of any given signal or peak depends on the structural environment surrounding the proton(s). The location of each peak is predictable, and the origins of the various signals or peaks within the spectrum have been fully characterized, facilitating the identification of the molecules within a specimen matrix. Fortuitously, the amplitudes of these signals are proportional to the number of molecules of the same kind that generated the signal. Using customized software, NMR signals can be

identified and signal amplitudes converted into concentrations, making NMR analysis of a specimen highly quantitative and reproducible. The strength of the magnetic field determines the resolution of the overlapping signals. For example, a 900-MHz magnet will provide the maximum resolution of the multitude of signals in a complex matrix. Even a 400-MHz proton NMR (^{1}H-NMR) spectrometer can quantify numerous, relevant analytes in common specimen matrices like serum, plasma, or urine with the highest accuracy and precision. Widely used in the clinical laboratory, 400- and 600-MHz ^{1}H-NMR instruments are relatively inexpensive. In contrast to other instrument systems commonly deployed throughout the laboratory, NMR spectrometers require minimal reagents and can simultaneously quantify many analytes with no preanalytical specimen manipulation. For these reasons, NMR spectroscopy has rapidly become a cost-effective means to measure analytes accurately and efficiently.

Nuclear magnetic resonance analyzers in clinical analysis

NMR analyzers are widely used in research and development, quality control, and quality assurance in food science, pharmaceutical, and chemical industries, as well as in academic, medical, and clinical arenas. They range in size, NMR type, resolution, and application, but have proven to be effective, powerful tools in their respective spectroscopic niches and disciplines.

Benchtop nuclear magnetic resonance spectrometers

A number of benchtop, low-field NMR analyzers exist on the market (Table 11.1). Distributed by different manufacturers, these analyzers have specific analytical chemistry applications for the identification of chemicals, compounds, and molecules in solid, semisolid, and liquid states. Routinely, these benchtop model NMR spectrometers are utilized in teaching laboratories and academic institutions for research purposes, performing a number of different types of experiments including, but not limited to, reaction monitoring and flow chemistry experiments. Similar to basic chemical reactions that are commonly monitored on large, automated chemistry platform in the laboratory today, these systems are capable of determining reaction end points and kinetics, as well as identification of intermediates or by-products. These benchtop systems also offer industrial applications for routine quality control and quality assurance of simple compounds to highly complex sample admixtures.

Of interest, and interrelated to pharmaceutical and clinical medicine, are applications for analysis and identification of prescribed and illicit drugs from seized samples as well as other biomarkers in biofluids [3,4]. For healthcare systems supporting large pain management populations, in partnership with law enforcement agencies and emergency room physicians, detection of active pharmaceutical ingredients (API) within seized drug samples is becoming increasingly important. Recent work using a high-throughput, benchtop NMR spectrometer and automated collection of ^{1}H-NMR spectra revealed that it is possible to easily screen for and detect novel psychoactive substances, herbal products mixed with synthetic cannabinoids, and illicit drugs. Spectral libraries from ^{1}H-NMR were compared with gas chromatography—mass spectrometry libraries as the currently accepted reference method, which demonstrated a 93% match of cases. NMR spectral analysis easily identified commonly abused drugs like cocaine, 3-4 methylenedioxymethamphetamine, heroin, mephedrone, ketamine, and diazepam in seized samples. This work confirms ^{1}H-NMR spectra generated via a high-throughput, benchtop NMR analytical platform that can rapidly ($\sim$5 minutes) identify API of drugs seized by law enforcement agencies [3]. Future work directed toward quantification methods to widen the applicability of qualitative identification may be of benefit to the clinical and forensic communities.

Other recent advancements in benchtop biofluid analysis have been made, specifically in urine and liquid biopsies [4]. Work in urinary biomarkers revealed that benchtop NMR systems successfully identified and quantified compounds like glucose and ketone bodies for diagnosis and monitoring of patients with type 2 diabetes (T2D). Benchtop 60-MHz ^{1}H-NMR quantitation of urinary glucose concentrations correlated well with chemical dipstick and glucose oxidase spectrophotometric methods. Interestingly, other biomarkers like alanine, citrate, and creatinine/creatine were found to be increased in this patient population, all consistent with previous reports of findings using high-field (400 or 600 MHz) ^{1}H-NMR instrument systems. Given the ease of specimen preparation, low cost, and increased investigations in biofluid applications, benchtop NMR systems are gaining appeal in near-patient, point-of-care fields [5]. Continued development and work using benchtop, low-field NMR systems may have future applications for spectroscopic analyses of a number of biomarkers in biological specimens, including blood, urine, saliva, and other body fluids.

Nuclear magnetic resonance expansion into the clinical laboratory

The first NMR analyzer to be cleared by the United States (US) Food and Drug Administration (FDA) for clinical diagnostic purposes was the 400 MHz Vantera Clinical Analyzer. The first clinical application for the

TABLE 11.1 Benchtop nuclear magnetic resonance spectrometers.

Manufacturer			
Thermo Fisher Scientific[a]	Magritek[b]	Nanalysis[c]	Oxford Instruments[d]
Models			
picoSpin 45 picoSpin 80	Spinsolve43 Spinsolve60 Spinsolve80 SpinsolveULTRA	NMReady60e NMReady60pro	MQC (F, 5, 23) Pulsar
NMR			
Proton	Boron Carbon Fluorine Nitrogen Phosphorus Proton	Boron Carbon Fluorine Lithium Phosphorus Proton	Carbon Fluorine Proton
Performance Larmor frequency (Proton NMR)			
45:45 ± 1 MHz 80:82 ± 2 MHz	42.5 MHz 60 MHz 80 MHz	60 MHz	MQC: 5–23 MHz Pulsar: 60 MHz
Sample volume requirements			
45:30 μL 80:40 μL	500 μL	Not given	200 μL–100 mL Model dependent
Applications			
Analytical chemistry Food science Petrochemical Pharmaceutical sciences Polymer research Reaction monitoring			

NMR, Nuclear magnetic resonance.

[a] *http://www.thermoscientific.com/en/products/nuclear-magnetic-resonance-nmr.html.*
[b] *http://www.magritek.com/.*
[c] *http://www.nanalysis.com/.*
[d] *http://www.oxford-instruments.com/products/spectrometers/nuclear-magnetic-resonance-nmr.*

Vantera was the NMR LipoProfile test, or lipoprotein profile, which was cleared by the FDA as an in vitro diagnostic test. The lipoprotein profile reports a number of total low-density lipoprotein particles (LDL-P) for physician use in the management of LDL-related cardiovascular disease (CVD) risk [6]. Results for several laboratory-developed tests (LDTs) are also included in the lipoprotein profile report, including the following: lipoprotein insulin resistance (LP-IR) index, which produces a score that varies from 0 to 100, with higher scores indicating insulin resistance; total high-density lipoprotein particles (HDL-P); small LDL particles; and LDL size. These analytes, however, represent only a fraction of the lipoprotein parameters that can be quantified using the proprietary software algorithms that report data for the lipoprotein profile. Jeyarajah and colleagues [7] described how the first software algorithm developed was able to quantitate all of the lipoprotein parameters from the signal envelope at ~0.8 ppm in the NMR spectrum (Fig. 11.2B). The broad peak, which arises from the methyl group protons of lipids within the lipoprotein particles, is deconstructed to provide concentrations for all lipoproteins based on their size, largely due to the fact that each lipoprotein of a given diameter emits its own distinct NMR signal (Fig. 11.2C). This phenomenon has been described in detail along with illustrations of the unique signals emitted by the lipoprotein subclasses (Fig. 11.2C) [7]. Specially designed software measures these distinct

signals and provides quantitative information regarding the lipoprotein particle classes [very-low-density lipoprotein (VLDL), LDL, and HDL] and subclasses [small, medium, and large], as well as weighted-average VLDL, LDL, and HDL sizes. In order to measure the individual lipoprotein parameters, the shape of the broad methyl signal is modeled as the sum of the lipoprotein subspecies signals using nonnegative linear least squares deconvolution [7]. This approach provides the signal area contributed by each lipoprotein subspecies, which after multiplication by their respective signal to particle conversion factors gives the subspecies particle concentrations. Subsequently, the concentration of each of the three lipoprotein classes [VLDL, LDL, and HDL] is determined by the sum of the individual particle subspecies. For example, total HDL particles are calculated by adding the concentrations of small, medium, and large HDL subclasses.

Since its introduction, the initial procedure for acquiring the 400-MHz proton NMR spectrum from a patient's serum or plasma sample has remained the same. However, the computational process by which lipoprotein subclass concentrations are derived by deconstructing the broad methyl signal has been optimized over the past 20 years. Optimization of the assay, represented by versions LP1, LP2, and LP3 of the proprietary software algorithm, has primarily involved expanding the number of subspecies in the mathematical model to better account for the compositional diversity of human lipoproteins. The most recently developed software algorithm, called LP4 (the fourth-generation lipoprotein profile algorithm), uses a further-optimized deconvolution method to quantify lipoprotein subspecies. One of the incentives for developing the new algorithm was to enable a more detailed investigation of the disease relationships of different HDL particle subspecies, prompted by emerging evidence for the functional and proteomic diversity of HDL particles.

The LP4 algorithm quantifies seven different HDL subspecies with good precision and corrects a prior overestimation of the absolute concentrations of HDL particles. The granularity in the number of HDL subspecies may provide additional means to understand the biological and clinical relevance of HDL, which is thought to have both antiinflammatory and antioxidant properties [8]. Also corrected was an aspect of prior mathematical modeling that led to an underestimation of LDL-P concentrations. There was no clinical consequence to this change, as both the previous LP3- and the new LP4-generated LDL-P concentrations show close associations with CVD outcomes [9]. Notably, the concentration of LDL particles generated by the LP4 algorithm is more closely related to the number of apolipoprotein B (apoB) molecules they carry. Moreover, corrections to the NMR-generated total HDL and LDL particle concentrations bring the NMR technique a step closer to standardization with other methods used in the clinical laboratory to estimate apolipoprotein and lipoprotein (e.g., HDL fractionation using the calibrated ion-mobility mass spectrometry) concentrations [10]. In all, the LP4 algorithm reports data for quite a few new parameters including the modified lipoprotein profile and several clinically relevant metabolites. Over the past 20 years, lipoprotein profiles have been determined for millions of patient samples as well as for samples from over 1500 observational and interventional clinical studies. Lipoprotein data from these studies have been used to show efficacy and safety in clinical studies of potential new therapeutics as well as to evaluate the potential clinical applications for NMR-measured analytes.

Clinical utility of lipoprotein particle numbers

Multiple lipoprotein parameters from the NMR lipoprotein profile have been shown to have clinical utility. LDL-P values are used by physicians for management of LDL, which is known to be one of the major contributors to, and risk factors for, CVD [6]. Traditionally, the cholesterol content of LDL (LDL-C) has been used to estimate circulating LDL concentrations. However, the LDL-C content of particles varies widely among individuals and is often dependent on the metabolic state of the subject whose sample is being evaluated [6]. For example, patients with insulin resistance have a higher number of small LDL and HDL particles containing less cholesterol and cholesteryl esters and some LDL particles that contain higher triglyceride content as compared with individuals without insulin resistance. Therefore in patients with insulin resistance, metabolic syndrome, and T2D, NMR-measured LDL-P is a better measure of LDL quantity as it is unaffected by changes in cholesteryl ester content and particle size [6]. In fact, discordance between LDL-C and LDL-P has been noted in several large clinical studies [11–15]. Notably, CVD risk tracks more closely with LDL-P than LDL-C [6]. Discordance between LDL-C and LDL-P has also been noted in statin-treated individuals. Consequently, multiple medical society expert panels have recommended the use of LDL-P as a target of therapy to guide medical decision-making with respect to management of LDL-related CVD risk, especially in patients with underlying metabolic disease [11–15]. The medical decision limit for LDL-P is >1000 nmol/L, and measurements above this threshold are associated with higher risk of CVD (Table 11.2) [15].

Besides LDL-P, NMR-measured HDL-P may be useful for assessment of residual CVD risk in primary or secondary prevention settings. Results from key epidemiological and clinical trials, including the Multi-Ethnic Study of Atherosclerosis (MESA), Heart Protection Study (HPS),

TABLE 11.2 Medical decision limits for NMR-measured CVD risk factors in adults (≥18 years).

Analyte (units)	Concentration	Classification
TC (mg/dL)	<200	Desirable
	200–239	Borderline high
	≥240	High
HDL-C (mg/dL)	<40	Low
	≥ 60	High
TG (mg/dL)	< 150	Normal
	150–199	Borderline high
	200–499	High
	≥500	Very high
ApoB (mg/dL)	<90	Low
	90–130	Intermediate
	>130	High
LDL-P (nmol/L)	<1000	Low
	1000–1299	Intermediate
	1300–1599	Borderline high
	≥1600	High
HDL-P (μmol/L)	<30	Low
	≥50	High
GlycA (μmol/L)	<400	Low
	≥400	High
TMAO (μmol/L)	<6.2	Low
	6.2–9.9	Intermediate
	≥10.0	High

ApoB, Apolipoprotein B; *HDL-C*, high-density lipoprotein cholesterol; *HDL-P*, high-density lipoprotein particles; *LDL-P*, low-density lipoprotein particles; *TC*, total cholesterol; *TG*, triglycerides; *TMAO*, trimethylamine-*N*-oxide.

and Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER), demonstrated that NMR-measured HDL-P was a stronger, more independent predictor of CVD risk than HDL cholesterol (HDL-C) content [8]. In JUPITER, investigators evaluated the relationship between HDL-C and HDL-P with CVD risk in more than 10,000 subjects randomized to receive rosuvastatin or placebo [16]. Similar to results from the HPS, JUPITER showed a strong, inverse association between the HDL-P concentrations and the CVD risk [16]. In contrast, HDL-C was not associated with CVD risk in statin-treated patients after adjustment for additional lipoprotein parameters [16]. Notably, with HDL-P measured by ion mobility, no association was observed with either incident CVD or CVD plus mortality in the JUPITER trial, suggesting that the HDL-P values obtained from NMR and ion mobility technologies are not equivalent. Therefore NMR-measured HDL-P has shown better promise as a marker of residual CVD risk than HDL-C and other measures of HDL-P [8]. The medical decision limit for HDL-P is <30 μmol/L, and values below this threshold are considered to be low and are associated with a higher risk of CVD (Table 11.2) [8].

Standard lipid panel and apolipoprotein B are measured by nuclear magnetic resonance

Recently, a traditional lipid panel assay, which reports results for total cholesterol (TC), triglycerides (TG), HDL-C and apoB, was added to the list of FDA-cleared tests on the Vantera Clinical Analyzer. This panel of traditional biomarkers has been used in patient care for many years, largely for the assessment of CVD risk. The NMR lipid panel results are substantially equivalent to the chemical measures for TC, TG, HDL-C, and apoB. While acquisition of the NMR spectra is the same as that used for the lipoprotein profile, the method for calculating the analyte concentrations is different. The software algorithm for determining results for the lipid panel and apoB uses a partial least squares (PLS) regression model that was trained on thousands of previously obtained NMR spectra, along with the chemistry results for TC, TG, HDL-C, and apoB. Using a PLS regression routine, the spectral information in the combined methylene and methyl region was trained against the chemical measurements. Cross-validation was performed to optimize the regression model. For any newly acquired NMR spectrum, the spectral information is then converted into TC, TG, HDL-C, or apoB concentrations. The use of PLS regression for the development of this assay differs from the traditional deconvolution technique used to quantify lipoprotein parameters. Precision for TC, TG, HDL-C, and apoB calculated from NMR spectra in this manner is similar to that reported by traditional chemical assays; % CV is for TC = 0.9%–1.6%, TG = 1.0%–1.4%, HDL-C = 1.3%–2.8%, and apoB = 1.2%–2.4%. Moreover, results from the NMR-measured lipid panel and apoB are comparable with the results from other assays vetted by the ReLABS Certification Program at the Northwest Lipid Research Laboratory at the University of Washington and the Lipid Standardization Program administered by the Center for Disease Control. This test increases the number of NMR assays that are available in the clinical laboratory for CVD risk assessment. Medical decision limits for the analytes of the standard lipid panel have been recommended by the The National Cholesterol

Education Program Adult Treatment Panel III Guidelines for the assessment and management of CVD risk (Table 11.2) [17–19]. Examples of reports and suggestions on how to interpret the results of these commonly reported laboratory tests can be found at the websites for various clinical laboratories.

LP-IR, a measure of insulin resistance that predicts future type 2 diabetes

As mentioned earlier, the LP-IR Index, which is currently an LDT, is useful for determination of a patient's insulin resistance status. Well-characterized changes in lipoprotein metabolism occur in subjects with insulin resistance, which may be observed in the NMR-measured lipoprotein parameters. Higher levels of large VLDL and small LDL particles and lower levels of large HDL particles have been observed [20]. In addition, mean VLDL particle size is generally greater, and mean LDL and HDL sizes are smaller, in insulin resistant subjects [20]. These six NMR-measured lipoprotein parameters have been combined into a single weighted LP-IR score (varying from 0 to 100); the higher the score, the more likely a patient is insulin-resistant [20]. Development of LP-IR was guided by data from the MESA cohort and verified by comparing LP-IR to glucose disposal rates measured during hyperinsulinemic-euglycemic clamps in insulin-sensitive, insulin-resistant, and untreated subjects with T2D [20]. The association of LP-IR with insulin resistance was validated in two independent cohorts of subjects studied at Duke University [21]. In both cohorts, there was a significant inverse correlation between LP-IR and the insulin sensitivity index measured during an intravenous glucose tolerance test, as well as a significant positive association with the homeostatic model assessment of insulin resistance (HOMA-IR) [21]. Taken together, these data suggest that LP-IR provides a simple, high-throughput means to identify individuals with insulin resistance from a single, fasting blood sample. Given that insulin resistance drives metabolic diseases such as T2D, knowing a patient's insulin resistance status is important for clinical decision-making and medical management.

Recently, it was shown in three varied study populations that LP-IR was able to predict future T2D [22–24]. In the MESA study, LP-IR was associated with future T2D with a hazard ratio (HR) of 1.59, even after adjusting for multiple diabetes risk factors including fasting blood glucose and TG to HDL-C ratio [23]. In the Women's Health Study (WHS), a study of nondiabetic women aged 45 or older, LP-IR was associated with incident T2D with a fully adjusted HR of 2.21 [22]. Similar to the MESA study, the WHS study showed that LP-IR was associated with T2D even in those subjects who would be considered to be of low risk. Hence, it was concluded that LP-IR detects insulin resistance and T2D risk years before clinically detectable dysglycemia. Moreover, in the WHS, LP-IR enhanced the performance of the Framingham Offspring Study diabetes risk score and gave improved performance when compared with 35 T2D risk factors, including a previously developed genetic risk score [22,25]. Furthermore, the JUPITER study showed that LP-IR was significantly associated with incident T2D in placebo and rosuvastatin-treated subjects after adjusting for traditional risk factors, body mass index (BMI), and lipids [24]. The authors concluded that the LP-IR score identifies individuals at risk for developing T2D including those taking statins, which have been shown in some studies to increase the risk of progressing to T2D.

In the Prevention of Renal and Vascular End Stage Disease (PREVEND) Study, a large cohort of men and women with a wide age range (28–75 years), LP-IR scores were strongly associated with incident T2D, even after adjustment for multiple T2D risk factors [26]. Moreover, LP-IR scores identified subjects with insulin resistance independently of BMI and glucose status. These results support the premise that LP-IR scores may be useful for identifying patients with insulin resistance who are at a risk of progressing to T2D, even when they do not show clinical signs of risk. While many of the multiparameter diabetes risk scores, such as those that are available for self-assessment on the American Diabetes Association website or based on the Framingham Offspring Study [25], include parameters that are not modifiable (e.g., age, gender, familial history of diabetes, and previous gestational diabetes), the LP-IR score takes into account lipoprotein changes that are fully modifiable with changes in diet and exercise [27–29]. Taken together, LP-IR may be a convenient way for physicians to assess insulin resistance and a patient's T2D risk, early in the progression of the disease, as well as monitor treatments aimed at reducing weight, improving insulin sensitivity, and preventing development of T2D. A medical decision limit or cut point for LP-IR and prediction of T2D was determined to be >68, above which a patient's risk of progressing to T2D would be high [26].

GlycA, a nuclear magnetic resonance-specific marker of systemic inflammation

GlycA, first described in 1987 and noted in several NMR metabolomics studies, is an example of the successful translation of an NMR-measured biomarker into a diagnostic tool for physician use [30]. The NMR resonance that is called GlycA was first noted by Bell et al. [31], who described a peak at approximately 2.04 ppm in the spectrum for human plasma was elevated in inflammation. The authors identified the signal as arising from the N-acetyl glucosamine residues located within the

carbohydrate side chains of acute phase proteins such as α1-acid glycoprotein, α1-antitrypsin, α1-antichymotrypsin, haptoglobin, and transferrin [30]. GlycA, quantified on the Vantera Clinical Analyzer as an LDT, has been demonstrated to be associated with common markers of inflammation such as high sensitivity C-reactive protein (hsCRP), fibrinogen, interleukin-6, and serum amyloid A [30–32]. Similar to what was found in the metabolomics studies, GlycA was shown to be higher in autoimmune diseases such as rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE) [31,33]. Moreover, GlycA was associated with disease activity and coronary artery atherosclerosis in patients with RA [31] as well as active inflammation at the vessel wall in patients with psoriasis [34]. Perhaps because it is a composite biomarker that measures both the increased protein levels and enhanced glycosylation states of the most abundant circulating acute phase proteins, GlycA has the advantage of having lower within-subject biological variation than hsCRP [30]. To illustrate this point, the homeostatic set points were determined for hsCRP and GlycA in a cohort of normal, healthy, adult volunteers collected for a previously published biological variability study [30]. This study calculated that 33 concurrent measurements of hsCRP would be required to estimate its true value for a single patient (similar to what was previously reported for hsCRP) compared with only one for GlycA, suggesting that GlycA is much less variable than hsCRP and, therefore, a more reliable measure of inflammation [30,35].

GlycA has been shown to be useful for predicting future CVD [36–38]. In the WHS, GlycA was associated with CVD events, independent of traditional risk factors [36]. In the PREVEND study, GlycA was associated with incident CVD, defined as the combined end point of CV morbidity and mortality, independent of clinical and lipid measures as well as renal function [37]. Baseline concentrations of GlycA in the JUPITER trial were significantly associated with incident CVD events, even when adjusting for established risk factors and a family history of premature coronary heart disease [38]. Remarkably, this association was only slightly attenuated by hsCRP [38]. GlycA was shown to be associated with future mortality [39,40]. GlycA was also found to be associated with CVD in patients undergoing angiography in the Intermountain Heart Collaborative Study [41]. In this large study, baseline levels of GlycA and hsCRP were the independent and additive markers of risk for future major adverse cardiac events, death, and hospitalization due to heart failure [41]. Given its usefulness for predicting future CVD and the fact that it is higher in RA, SLE, and psoriasis, GlycA may be useful for CVD risk assessment in patients with autoimmune diseases, where traditional CVD risk factors such as LDL-C and hsCRP are not predictive of CVD [42]. The medical decision limit for GlycA is ≥ 400 μmol/L, above which a patient would be considered to have high systemic inflammation, have higher CVD risk, and may warrant clinical management (Table 11.2) [30].

Clinical utility for nuclear magnetic resonance quantified metabolites

NMR has been used for 20 years for quantifying lipoprotein particles; however, only recently has significant attention been given to the wealth of additional analytes that can be quantified from the typical NMR spectrum. Metabolic phenotyping by NMR has become a widely used tool to identify analytes that have significant associations with various disease states [43,44]. Metabolomics, applied in the academic setting to many epidemiological studies, has uncovered NMR-measured analytes that are associated with metabolic disorders such as insulin resistance, metabolic syndrome, T2D, and diabetic complications [43]. These publications strengthen the already mounting evidence that suggests that there are predictable alterations in fatty acid, amino acid, and lipoprotein metabolism that occur in various metabolic diseases. Circulating fatty acids, TG, VLDL, and LDL particles, VLDL particle size, branched-chain and aromatic amino acids, and markers of inflammation are higher, whereas HDL particles, HDL and LDL particle size, adiponectin, glutamine, and β-hydroxybutyrate (β-HB) are lower in subjects with higher adiposity, metabolic syndrome, and T2D [45,46]. Branched-chain amino acids (BCAA) have been found to be altered in metabolomics studies of alcoholic liver disease and nonalcoholic fatty (NAFLD) liver disease, while NMR-measured N-acetyl glycoproteins, LDL, creatine, and albumin were decreased and citrate and VLDL were increased in subjects with chronic hepatitis C-induced liver fibrosis [47–49]. NMR metabolomics studies have also been applied in the area of CVD diagnostics in order to enhance CVD risk assessment [2,43,50,51]. Associations between NMR-measured analytes and mortality as well as Alzheimer's disease have also been uncovered [39,52]. Finally, metabolomics studies have been employed for the purpose of discovering novel biomarkers that might be useful for diagnosis, prognosis, or monitoring of disease activity for autoimmune diseases such as RA, SLE, and inflammatory bowel disease [53,54].

While metabolomics studies are useful for uncovering disease associations, in order to apply the metabolomics information for clinical diagnostic purposes, software algorithms should be developed that measure the analyte of interest in an absolute sense. These algorithms should, therefore, take into account the many overlapping signals that may occur in the same region as the desired analyte. For example, the NMR spectrum from a patient with severe hypertriglyceridemia will look very different from

that of a patient with a normal lipid profile. The NMR signals from the TG molecules overlap several key analytes and will have profound effects on their quantitation if not considered during the deconvolution process. This indicates that NMR, like other general chemistry and immunoassay platforms, is not free from common interferences and requires development and validation of appropriate deconvolutions for accurate analyte quantitation. Once complete, the assays may then be used to interrogate a patient's sample with confidence that the absolute concentration of the desired analyte has been determined despite the potential for interfering peaks that may be unique to their profile.

Branched-chain amino acids (BCAA) for prediction of T2D risk

The BCAA (valine, leucine, and isoleucine) are essential amino acids that are not only required for protein synthesis but also have been shown to regulate protein production, protein degradation, and glucose metabolism. Decreased circulating concentrations of BCAA are associated with several pathological states including liver disease, early chronic kidney disease, and all-cause mortality [55]. Increased BCAA, on the other hand, are associated with insulin resistance, T2D, coronary artery disease, and an altered microbiome [55]. While it has been hypothesized that increased circulation of BCAA may be a causal factor in the development of insulin resistance and T2D, recent genetic evidence supports the concept that insulin resistance may in fact drive higher circulating fasting BCAA concentrations [55]. BCAA can be measured from the same NMR spectra acquired for the lipoprotein profile by taking into account the overlapping signals from the proteins and lipoproteins in each sample, thereby allowing quantification from spectra collected for multiple clinical purposes [6,30]. Similar to the NMR LipoProfile and GlycA tests, the BCAA assay is high throughput with an analytical time of 90 seconds and there is no requirement for sample preparation before placing it on the clinical instrument for analysis. Furthermore, the analytical performance data revealed that the BCAA assay is robust and suitable for clinical testing [55].

BCAA are predictive of T2D development and are responsive to therapeutic interventions that enhance insulin sensitivity [55]. In fact, fasting concentrations of BCAA were found to be elevated before the onset of T2D [56]. Consistent with recent literature, analysis from two independent cohorts showed that NMR-measured BCAA are elevated in subjects with T2D compared with nondiabetic subjects [55]. Furthermore, BCAA demonstrated independent and positive association with T2D and metabolic syndrome in each cohort, and in subjects with varying degrees of glucose tolerance, NMR-determined BCAA concentrations have been shown to closely relate to insulin resistance, as assessed by HOMA-IR [57]. Observed in a small number of participants, this study also suggested that BCAA concentrations may be directly related to carotid intima-media thickness, a proxy of subclinical atherosclerosis, and independent of the presence of T2D and metabolic syndrome [58–61]. Though preliminary, this cross-sectional association added to recent observations showing that BCAA concentrations may predict incident CVD [62]. BCAA measurements may be useful for determining early metabolic dysfunction, well in advance of the development of chronic disease, thereby enabling preventative measures to reduce potentially disease progression. Medical decision limits for BCAA have not yet been established.

Ketone bodies for diagnosis of diabetic or alcoholic ketoacidosis

Ketone bodies (acetone, acetoacetic acid, beta-hydroxybutyric acid) are products of fat catabolism that are used as alternative sources of energy to glucose when carbohydrate intake is low and there is a surplus of circulating free fatty acids [63]. Elevations in ketone bodies occur when there is insufficient insulin to utilize glucose as an energy source (ketoacidosis), during times of prolonged fasting (mild ketonemia) or when consuming a carbohydrate-restricted diet (ketogenic diet). The predominant ketones are β-HB, acetoacetate, and acetone. Under normal physiological conditions, total plasma ketone concentrations fluctuate between 100 and 600 μM, but can rise to $\sim$1 mM after prolonged exercise or 24 hours of fasting [63]. For example, β-HB concentrations $>$2.8 mg/dL ($>$1.0–1.5 mM) are indicative of ketoacidosis and will alert a physician to the need for immediate treatment and follow-up testing. Concentrations of the three ketone bodies combined can rise as high as 6–20 mM in patients with diabetic or alcoholic ketoacidosis [64,65]. Therefore ketone bodies are key for diagnosis of diabetic and alcoholic ketoacidosis, both of which can be fatal if not recognized before symptoms become life threatening [66,67]. Like the BCAA, the three main ketone bodies can be concurrently quantified from lipoprotein profile spectra [66,67].

While measurement of ketone body concentrations is critical to diagnosis of ketoacidosis, it may also be useful for quantifying circulating ketones when there is only a mild elevation. For example, diets that are rich in proteins and fats but very low in carbohydrates (e.g., Atkins diet; also known as a ketogenic diet) have been shown to induce mild elevations of ketone bodies. Elevations in ketones produced while consuming a ketogenic diet are much lower than those observed in ketoacidosis and are not associated with a change in blood pH [68]. However,

since very high ketone levels may have nonbeneficial effects, physicians or patients may want to monitor ketone bodies to ensure the diet is managed correctly and is beneficial [69]. In addition, sodium-glucose transporter-2 (SGLT2) inhibitor treatment for T2D may elicit mild ketonemia that is reminiscent of the mild elevations induced by ketogenic diets and may be related to their ability to reduce plasma glucose via renal elimination, stimulate lipolysis, elicit weight loss, and enhance insulin sensitivity [70,71]. Physicians may, therefore, wish to monitor their patients to ensure that underlying pathologies do not contribute to higher than usual increases in ketone bodies while consuming a ketogenic diet or on SGLT2 inhibitor therapy.

Trimethylamine-N-oxide for assessment of gut dysfunction and cardiovascular disease risk

Trimethylamine-N-Oxide (TMAO) is a metabolite that is increasingly recognized as clinically important in the assessment of CVD [72]. TMAO is produced by a pathway whose initial rate-limiting step is gut microbe-dependent [72]. It has been observed that gut bacteria are responsible for metabolizing dietary choline and carnitine into trimethylamine, which is then transformed into TMAO in the liver. A study where oral broad-spectrum antibiotics were used to suppress the intestinal microbiota led to a reduction in circulating TMAO levels, providing direct evidence that dietary status and the gut microbiome affect TMAO concentrations [72]. While BCAA and ketone bodies are quantified from the same NMR spectrum as the lipoprotein profile, the TMAO LDT involves the acquisition of a proton Carr−Purcell−Meiboom−Gill (CPMG) NMR spectrum from a plasma or serum sample that has been adjusted to pH 5.3. The acquisition of the CPMG spectrum allows for suppression of signals from lipoproteins and proteins whose signals interfere with quantification of the signal peaks from small molecule metabolites. Moreover, the adjustment to pH 5.3 shifts the TMAO signal to a less crowded place in the spectrum (3.29−3.30 ppm), where it can be readily quantified without interfering signals from metabolites like betaine. The NMR signal is then modeled with a proprietary fitting algorithm.

Several clinical studies have shown an association of elevated plasma TMAO concentrations with increased risk of CVD, independent of established risk factors. A study of adults undergoing cardiac catheterization revealed that subjects with high TMAO had a 2.5-fold increased risk for a major adverse CVD event (myocardial infarction, stroke, or death) compared with subjects in the lowest quartile [73]. Subjects with underlying chronic kidney disease had higher fasting plasma TMAO levels and a higher risk of mortality than those with normal renal function [74]. TMAO levels showed a significant, graded association with CVD but not with measured carotid intima-media thickness [75]. In a cohort of patients with a history of heart failure, elevated plasma TMAO was associated with increased risk of mortality, independent of traditional risk factors and B-type natriuretic peptide levels. It has been shown that subjects with cardiometabolic diseases such as T2D and NAFLD exhibit altered gut microbiome and higher circulating levels of TMAO compared with healthy subjects or subjects on a vegan diet. In fact, a short-term, high-fat diet alone increased postprandial levels of circulating TMAO. Current thought in the field suggests that the bacterial changes and accompanying alterations in metabolic signaling that are observed in the gut microbiota of subjects with metabolic diseases elicit increases in TMAO, which then contributes to increased CVD risk in these patients. It has been hypothesized that changes in diet or dietary supplementation may reduce metabolic disease and its associated CVD risk, in part by reducing circulating levels of TMAO. Therefore monitoring TMAO levels before and after changes in diet and/or treatment with pre- or probiotics or dietary supplements may help physicians encourage patients to change their lifestyle and reduce long-term risk of CVD events. The medical decision limits for TMAO with respect to CVD risk are as follows: <6.2 low risk, 6.2−9.9 moderate or intermediate risk, and ≥ 10.0 for high risk (Table 11.2) [72].

Nuclear magnetic resonance instrument and diagnostic test manufacturers entering the US diagnostic market

Emerging into the US NMR-based diagnostic care market, Numares Health, based in Germany, is one of the new players in the NMR-based diagnostic business. The company has marketed the Axinon lipoFIT assay, which is deployed on the Axinon 600-MHz NMR system [76−78]. Similar to the NMR LipoProfile Test, the lipoFIT test reports data for the standard lipid panel (TC, TG, LDL-C, and HDL-C) as well as for key lipoprotein particle concentrations (LDL-P, HDL-P, large VLDL-P, small LDL-P, large HDL-P, VLDL, LDL, and HDL size). This test is currently a research use only (RUO) product in the United States but holds a CE Mark in the European Union market. Recently, Numares Health brought forward another RUO test called the Axinon renalTX-score. This test is a metabolomics-based urine test for detection of kidney allograft rejection. This test is based on the results of four metabolites (alanine, lactate, urea, and citrate) and provides a score (0−100); the higher the score, the higher the probability of acute allograft rejection.

Like Numares Health, Bruker, also based in Germany, has recently entered the US market with an RUO test for

lipoprotein subclass analysis. This test, which is run on an Avance 600-MHz NMR instrument, reports results for the standard lipid panel (TC, TG, LDL-C, and HDL-C) as well as for key apolipoprotein (apoA-I, apoA-II, and apoB) and lipoprotein particle concentrations. Finally, Biosfer Teslab, based in Spain, has developed the LipoScale test for lipoprotein analysis as well as a test to measure GlycA [79,80]. A review of the various software techniques used to quantify lipoprotein parameters by these companies was published by Mallol et al. [81]. The number of available NMR tests and platforms is growing and may lead to new NMR-based diagnostic tests in the coming years.

Conclusions

This chapter provided an overview of theory, principles, and clinical applications of NMR technology. From bench to bedside, advancements in the NMR field have transformed how scientists, and clinical and laboratory professionals interrogate the human body—both physically, as images, and biochemically. Research and spectroscopic interrogation of biological specimens from normal populations as well as from a number of disease states using low- and high-field NMR will continue to revolutionize clinical applications of NMR in laboratory medicine.

Disclosure

Dr. Connelly is an employee of LabCorp, the company that owns the patents and technology for the Vantera Clinical Analyzer.

References

[1] E.D.F.C. Becker, C.L. Khetrapal, The development of NMR, in: D.M. Grant, R.K. Harris (Eds.), Encyclopedia of Nuclear Magnetic Resonance, Wiley, Chichester, 1996, pp. 1–160.

[2] N.J. Rankin, D. Preiss, P. Welsh, et al., The emergence of proton nuclear magnetic resonance metabolomics in the cardiovascular arena as viewed from a clinical perspective, Atherosclerosis 237 (1) (2014) 287–300.

[3] L.H. Antonides, R.M. Brignall, A. Costello, et al., Rapid identification of novel psychoactive and other controlled substances using low-field (1)H NMR spectroscopy, ACS Omega 4 (4) (2019) 7103–7112.

[4] M. Grootveld, B. Percival, M. Gibson, et al., Progress in low-field benchtop NMR spectroscopy in chemical and biochemical analysis, Analytic. Chim. Acta 1067 (2019) 11–30.

[5] B.C. Percival, M. Grootveld, M. Gibson, et al., Low-field, benchtop NMR spectroscopy as a potential tool for point-of-care diagnostics of metabolic conditions: validation, protocols and computational models, High. Throughput 8 (1) (2018).

[6] S.P. Matyus, P.J. Braun, J. Wolak-Dinsmore, et al., NMR measurement of LDL particle number using the Vantera clinical analyzer, Clin. Biochem. 47 (16–17) (2014) 203–210.

[7] E.J. Jeyarajah, W.C. Cromwell, J.D. Otvos, Lipoprotein particle analysis by nuclear magnetic resonance spectroscopy, Clin. Lab. Med. 26 (4) (2006) 847–870.

[8] S.P. Matyus, P.J. Braun, J. Wolak-Dinsmore, et al., HDL particle number measured on the Vantera(R), the first clinical NMR analyzer, Clin. Biochem. 48 (3) (2015) 148–155.

[9] M.H. Davidson, C.M. Ballantyne, T.A. Jacobson, et al., Clinical utility of inflammatory markers and advanced lipoprotein testing: advice from an expert panel of lipid specialists, J. Clin. Lipidol. 5 (5) (2011) 338–367.

[10] P.M. Hutchins, G.E. Ronsein, J.S. Monette, et al., Quantification of HDL particle concentration by calibrated ion mobility analysis, Clin. Chem. 60 (11) (2014) 1393–1401.

[11] A.J. Garber, M.J. Abrahamson, J.I. Barzilay, et al., AACE comprehensive diabetes management algorithm 2013, Endocr. Pract 19 (2) (2013) 327–336.

[12] T.A. Jacobson, M.K. Ito, K.C. Maki, et al., National lipid association recommendations for patient-centered management of dyslipidemia: Part 1—Full report, J. Clin. Lipidol. 9 (2) (2015) 129–169.

[13] P.S. Jellinger, Y. Handelsman, P.D. Rosenblit, et al., American Association of Clinical Endocrinologists and American College of Endocrinology Guidelines for Management of Dyslipidemia and Prevention of Cardiovascular Disease, Endocr. Pract 23 (Suppl. 2) (2017) 1–87.

[14] A.J. Garber, M.J. Abrahamson, J.I. Barzilay, et al., Consensus statement by the American Association of Clinical Endocrinologists and American College of Endocrinology on the comprehensive type 2 diabetes management algorithm - 2019 executive summary, Endocr. Pract. 25 (1) (2019) 69–100.

[15] W.C.T.D.W. Cromwell, Clinical utility of LDL particle number to optimize management of LDL-related cardiovascular risk, J. Family Pract. 65 (7) (2016) 1–8.

[16] S. Mora, R.J. Glynn, P.M. Ridker, High-density lipoprotein cholesterol, size, particle number, and residual vascular risk after potent statin therapy, Circulation 128 (11) (2013) 1189–1197.

[17] DeWitt S. Goodman, MD; Stephen B. Hulley, MD, MPH; Luther T. Clark, MD; et al., Report of the National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. The expert panel. Arch. Intern. Med. 148 (1) (1988) 36–69.

[18] Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, Executive summary of the third report of The National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III), JAMA 285 (19) (2001) 2486–2497.

[19] S.M. Grundy, Low-density lipoprotein, non-high-density lipoprotein, and apolipoprotein B as targets of lipid-lowering therapy, Circulation 106 (20) (2002) 2526–2529.

[20] I. Shalaurova, M.A. Connelly, W.T. Garvey, J.D. Otvos, Lipoprotein insulin resistance index: a lipoprotein particle-derived measure of insulin resistance, Metab. Syndr. Relat. Disord. 12 (8) (2014) 422–429.

[21] C.R. Mikus, M.J. Patel, C.A. Slentz, L.A. Baterman, L.H. Willis, W.E. Kraus, Comparison of lipoprotein-derived insulin resistance score and standard measures of insulin sensitivity, Diabetes 62 (2013). A99-!172.

[22] P.H.N. Harada, O.V. Demler, S.B. Dugani, et al., Lipoprotein insulin resistance score and risk of incident diabetes during extended follow-up of 20 years: the women's health study, J. Clin. Lipidol. 11 (5) (2017) 1257–1267e1252.

[23] R.H. Mackey, S. Mora, A.G. Bertoni, et al., Lipoprotein Particles and Incident Type 2 Diabetes in the Multi-Ethnic Study of Atherosclerosis, Diabetes Care (2015).
[24] A.O. Dugani SBA, R.J. Glynn, P.M. Ridker, S. Mora, Lipoprotein particle characteristics and incident diabetes in the JUPITERTrial, Circulation 130 (Suppl. 2) (2014) A16055.
[25] P.W. Wilson, J.B. Meigs, L. Sullivan, C.S. Fox, D.M. Nathan, R. B. D'Agostino Sr., Prediction of incident diabetes mellitus in middle-aged adults: the Framingham offspring study, Arch. Intern. Med. 167 (10) (2007) 1068–1074.
[26] J.L. Flores-Guerrero, M.A. Connelly, I. Shalaurova, et al., Lipoprotein insulin resistance index, a high-throughput measure of insulin resistance, is associated with incident type II diabetes mellitus in the prevention of renal and vascular end-stage disease study, J. Clin. Lipidol. 13 (1) (2019) 129–137e121.
[27] D.L. Ellsworth, N.S. Costantino, H.L. Blackburn, R.J. Engler, M. Kashani, M.N. Vernalis, Lifestyle modification interventions differing in intensity and dietary stringency improve insulin resistance through changes in lipoprotein profiles, Obes. Sci. Pract. 2 (3) (2016) 282–292.
[28] S. Fernandez-Castillejo, R.M. Valls, O. Castaner, et al., Polyphenol rich olive oils improve lipoprotein particle atherogenic ratios and subclasses profile: a randomized, crossover, controlled trial, Mol. Nutr. Food Res. 60 (7) (2016) 1544–1554.
[29] N.H. Bhanpuri, S.J. Hallberg, P.T. Williams, et al., Cardiovascular disease risk factor responses to a type 2 diabetes care model including nutritional ketosis induced by sustained carbohydrate restriction at 1 year: an open label, non-randomized, controlled study, Cardiovasc. Diabetol. 17 (1) (2018) 56.
[30] J.D. Otvos, I. Shalaurova, J. Wolak-Dinsmore, et al., GlycA: a composite nuclear magnetic resonance biomarker of systemic inflammation, Clin. Chem. 61 (5) (2015) 714–723.
[31] J.D. Bell, J.C. Brown, J.K. Nicholson, P.J. Sadler, Assignment of resonances for 'acute-phase' glycoproteins in high resolution proton NMR spectra of human blood plasma, FEBS Lett 215 (1987) 311–315.
[32] R.P. Dullaart, E.G. Gruppen, M.A. Connelly, J.D. Lefrandt, A pro-inflammatory glycoprotein biomarker is associated with lower bilirubin in metabolic syndrome, Clin. Biochem. (2015).
[33] C.P. Chung, M.J. Ormseth, M.A. Connelly, et al., GlycA, a novel marker of inflammation, is elevated in systemic lupus erythematosus, Lupus 25 (3) (2016) 296–300.
[34] A.A. Joshi, J.B. Lerman, T.M. Aberra, et al., GlycA is a novel biomarker of inflammation and subclinical cardiovascular disease in psoriasis, Circulation Res. 119 (11) (2016) 1242–1253.
[35] C.G. Fraser, Test result variation and the quality of evidence-based clinical guidelines, Clin. Chim. Acta. 346 (1) (2004) 19–24.
[36] A.O. Akinkuolie, J.E. Buring, P.M. Ridker, S. Mora, A novel protein glycan biomarker and future cardiovascular disease events, J. Am. Heart Assoc. 3 (5) (2014) e001221.
[37] E.G. Gruppen, I.J. Riphagen, M.A. Connelly, J.D. Otvos, S.J. Bakker, R.P. Dullaart, GlycA, a pro-inflammatory glycoprotein biomarker, and incident cardiovascular disease: relationship with C-reactive protein and renal function, PLoS One 10 (9) (2015) e0139057.
[38] A.O.G.R.J. Akinkuolie, P.M. Ridker, S. Mora, Protein glycan side-chains, rosuvastatin therapy, and incident vascular events: an anlysis from the JUPITER trial, Circulation 130 (Suppl. 2) (2014) A17731.
[39] K. Fischer, J. Kettunen, P. Wurtz, et al., Biomarker profiling by nuclear magnetic resonance spectroscopy for the prediction of all-cause mortality: an observational study of 17,345 persons, PLoS Med. 11 (2) (2014) e1001606.
[40] P. Ritchie SCW, A.P. Nath, G. Abraham, et al., Systems medicine links microbial inflammatory response with glycoprotein-associated mortality risk, bioRxiv (2015).
[41] J.B. Muhlestein, H.T. May, O. Galenko, et al., GlycA and hsCRP are independent and additive predictors of future cardiovascular events among patients undergoing angiography: the intermountain heart collaborative study, Am. Heart J. 202 (2018) 27–32.
[42] A. Bag-Ozbek, J.T. Giles, Inflammation, adiposity, and atherogenic dyslipidemia in rheumatoid arthritis: is there a paradoxical relationship? Curr. Allergy Asthm. Rep. 15 (2) (2015) 497.
[43] P. Soininen, A.J. Kangas, P. Wurtz, T. Suna, M. Ala-Korpela, Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics, Circulation Cardiovasc. Genet. 8 (1) (2015) 192–206.
[44] M. Ala-Korpela, A.J. Kangas, P. Soininen, Quantitative high-throughput metabolomics: a new era in epidemiology and genetics, Genome Med. 4 (4) (2012) 36.
[45] P. Wurtz, Q. Wang, A.J. Kangas, et al., Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change, PLoS Med. 11 (12) (2014) e1001765.
[46] M. Fizelova, M. Miilunpohja, A.J. Kangas, et al., Associations of multiple lipoprotein and apolipoprotein measures with worsening of glycemia and incident type 2 diabetes in 6607 non-diabetic Finnish men, Atherosclerosis 240 (1) (2015) 272–277.
[47] W. Zhong, Z. Zhou, Alterations of the gut microbiome and metabolome in alcoholic liver disease, World J. Gastrointest. Pathophysiol. 5 (4) (2014) 514–522.
[48] S. Cheng, P. Wiklund, R. Autio, et al., Adipose tissue dysfunction and altered systemic amino acid metabolism are associated with non-alcoholic fatty liver disease, PLoS One 10 (10) (2015) e0138889.
[49] C.J. Sands, I.N. Guha, M. Kyriakides, et al., Metabolic phenotyping for enhanced mechanistic stratification of chronic hepatitis C-induced liver fibrosis, Am. J. Gastroenterol. 110 (1) (2015) 159–169.
[50] I.E. Hoefer, S. Steffens, M. Ala-Korpela, et al., Novel methodologies for biomarker discovery in atherosclerosis, Eur. Heart J. 36 (39) (2015) 2635–2642.
[51] P. Wurtz, A.S. Havulinna, P. Soininen, et al., Metabolite profiling and cardiovascular event risk: a prospective study of three population-based cohorts, Circulation (2015).
[52] T. Tukiainen, P. Jylanki, V.P. Makinen, et al., Mild cognitive impairment associates with concurrent decreases in serum cholesterol and cholesterol-related lipoprotein subclasses, J. Nutr. Health Aging 16 (7) (2012) 631–635.
[53] M.B. Lauridsen, H. Bliddal, R. Christensen, et al., 1H NMR spectroscopy-based interventional metabolic phenotyping: a cohort study of rheumatoid arthritis patients, J. Proteome Res. 9 (9) (2010) 4545–4553.
[54] R. Schicho, R. Shaykhutdinov, J. Ngo, et al., Quantitative metabolomic profiling of serum, plasma, and urine by (1)H NMR spectroscopy discriminates between patients with inflammatory bowel

disease and healthy individuals, J. Proteome Res. 11 (6) (2012) 3344−3357.
[55] J. Wolak-Dinsmore, E.G. Gruppen, I. Shalaurova, et al., A novel NMR-based assay to measure circulating concentrations of branched-chain amino acids: elevation in subjects with type 2 diabetes mellitus and association with carotid intima media thickness, Clin. Biochem. 54 (2018) 92−99.
[56] T.J. Wang, M.G. Larson, R.S. Vasan, et al., Metabolite profiles and the risk of developing diabetes, Nat. Med. 17 (4) (2011) 448−453.
[57] M.A. Connelly, J. Wolak-Dinsmore, R.P.F. Dullaart, Branched chain amino acids are associated with insulin resistance independent of leptin and adiponectin in subjects with varying degrees of glucose tolerance, Metab. Syndr. Relat. Disord. 15 (4) (2017) 183−186.
[58] S.H. Shah, J.L. Sun, R.D. Stevens, et al., Baseline metabolomic profiles predict cardiovascular events in patients at risk for coronary artery disease, Am. Heart J. 163 (5) (2012) 844−850e841.
[59] C.J. Lynch, S.H. Adams, Branched-chain amino acids in metabolic signalling and insulin resistance, Nat. Rev. Endocrinol. 10 (12) (2014) 723−736.
[60] S. Bhattacharya, C.B. Granger, D. Craig, et al., Validation of the association between a branched chain amino acid metabolite profile and extremes of coronary artery disease in patients referred for cardiac catheterization, Atherosclerosis 232 (1) (2014) 191−196.
[61] M. Ruiz-Canela, E. Toledo, C.B. Clish, et al., Plasma branched-chain amino acids and incident cardiovascular disease in the PREDIMED trial, Clin. Chem. 62 (4) (2016) 582−592.
[62] D.K. Tobias, S. Mora, S. Verma, P.R. Lawler, Altered branched chain amino acid metabolism: toward a unifying cardiometabolic hypothesis, Curr. Opcardiol 33 (5) (2018) 558−564.
[63] L. Laffel, Ketone bodies: a review of physiology, pathophysiology and application of monitoring to diabetes, Diabetes Metab. Res. Rev. 15 (6) (1999) 412−426.
[64] D.S. Schade, R.P. Eaton, Differential diagnosis and therapy of hyperketonemic state, JAMA. 241 (19) (1979) 2064−2065.
[65] R. Bienia, I. Ripoll, Diabetic ketoacidosis, JAMA. 241 (5) (1979) 510−511.
[66] A. Pinto, A. Bonucci, E. Maggi, M. Corsi, R. Businaro, Anti-oxidant and anti-inflammatory activity of ketogenic diet: new perspectives for neuroprotection in alzheimer's disease, Antioxid. (Basel) 7 (5) (2018).
[67] H. Chandrasekara, P. Fernando, M. Danjuma, C. Jayawarna, Ketoacidosis is not always due to diabetes, BMJ Case Rep. 2014 (2014).
[68] V.J. Miller, F.A. Villamena, J.S. Volek, Nutritional ketosis and mitohormesis: potential implications for mitochondrial function and human health, J. Nutr. Metab. 2018 (2018) 5157645.
[69] L. Gupta, D. Khandelwal, S. Kalra, P. Gupta, D. Dutta, S. Aggarwal, Ketogenic diet in endocrine disorders: current perspectives, J. Postgrad. Med. 63 (4) (2017) 242−251.
[70] D. Polidori, H. Iijima, M. Goda, N. Maruyama, N. Inagaki, P.A. Crawford, Intra- and inter-subject variability for increases in serum ketone bodies in patients with type 2 diabetes treated with the sodium glucose co-transporter 2 inhibitor canagliflozin, Diabetes Obes. Metab. 20 (5) (2018) 1321−1326.
[71] S.H. Min, T.J. Oh, S.I. Baek, et al., Degree of ketonaemia and its association with insulin resistance after dapagliflozin treatment in type 2 diabetes, Diabetes Metab. 44 (1) (2018) 73−76.
[72] E. Garcia, J. Wolak-Dinsmore, Z. Wang, et al., NMR quantification of trimethylamine-N-oxide in human serum and plasma in the clinical laboratory setting, Clin. Biochem. 50 (16−17) (2017) 947−955.
[73] Z. Wang, E. Klipfell, B.J. Bennett, et al., Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease, Nature 472 (7341) (2011) 57−63.
[74] W.H. Tang, Z. Wang, D.J. Kennedy, et al., Gut microbiota-dependent trimethylamine N-oxide (TMAO) pathway contributes to both development of renal insufficiency and mortality risk in chronic kidney disease, Circ. Res. 116 (3) (2015) 448−455.
[75] W.H. Tang, Z. Wang, Y. Fan, et al., Prognostic value of elevated levels of intestinal microbe-generated metabolite trimethylamine-N-oxide in patients with heart failure: refining the gut hypothesis, J. Am. Coll. Cardiol. 64 (18) (2014) 1908−1914.
[76] A. Vehtari, V.P. Makinen, P. Soininen, et al., A novel Bayesian approach to quantify clinical variables and to determine their spectroscopic counterparts in 1H NMR metabonomic data, BMC Bioinformatics. 8 (Suppl. 2) (2007) S8.
[77] B. Kaess, M. Fischer, A. Baessler, et al., The lipoprotein subfraction profile: heritability and identification of quantitative trait loci, J. Lipid Res. 49 (4) (2008) 715−723.
[78] A.K. Petersen, K. Stark, M.D. Musameh, et al., Genetic associations with lipoprotein subfractions provide information on their biological nature, Hum. Mol. Genet. 21 (6) (2012) 1433−1443.
[79] R. Mallol, N. Amigo, M.A. Rodriguez, et al., Liposcale: a novel advanced lipoprotein test based on 2D diffusion-ordered 1H NMR spectroscopy, J. Lipid Res. 56 (3) (2015) 737−746.
[80] R. Fuertes-Martin, D. Taverner, J.C. Vallve, et al., Characterization of (1)H NMR plasma glycoproteins as a new strategy to identify inflammatory patterns in rheumatoid arthritis, J. Proteome Res. 17 (11) (2018) 3730−3739.
[81] R. Mallol, M.A. Rodriguez, J. Brezmes, L. Masana, X. Correig, Human serum/plasma lipoprotein analysis by NMR: application to the study of diabetic dyslipidemia, Prog. Nucl. Magn. Reson. Spectrosc. 70 (2013) 1−24.

Self-assessment questions

1. Nuclei are "in resonance" with an external magnetic field when they are:
   - **a.** aligned in the opposite direction as the magnetic field.
   - **b.** aligned in the same direction as the magnetic field.
   - **c.** randomly flipping in the magnetic field.
2. Clinical applications of NMR include the quantitative determination of:
   - **a.** lipoproteins
   - **b.** GlycA
   - **c.** LP-IR
   - **d.** ketone bodies
   - **e.** all of the above
3. Which isotopes have "NMR active" nuclei?
   - **a.** carbon, fluorine, hydrogen, nitrogen, and phosphorus
   - **b.** calcium, iron, magnesium, and potassium
   - **c.** chloride, potassium, and sodium
4. A Vantera Clinical Analyzer with a _____ MHz magnet was the first NMR analyzer to be cleared by the FDA for use in the clinical laboratory.
   - **a.** 40
   - **b.** 400
   - **c.** 600
   - **d.** 900
5. Which NMR-measured metabolites have clinical utility?
   - **a.** BCAA
   - **b.** ketone bodies
   - **c.** TMAO
   - **d.** all of the above
6. LDL particle number of LDL-P can be used by physicians:
   - **a.** for the management of a patient's LDL-related CVD risk.
   - **b.** to diagnose RA.
   - **c.** to monitor the gut microbiome.

Answers

1. b
2. e
3. a
4. b
5. d
6. a

Chapter 12

Immunoassays

William Clarke[1], **Lori J. Sokoll**[1] **and Alex J. Rai**[2]

[1]Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD, United States, [2]Department of Pathology, Columbia University Medical Center, New York, NY, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- List factors that affect antibody–antigen binding
- Discuss the various immunoassay formats used in clinical laboratories
- Explain the advantages and disadvantages of using immunoassays for quantitative analyses
- Discuss potential interferences with clinical immunoassays
- Discuss the use of immunoassays in point-of-care testing

Introduction

Immunoassays are analytical methods based on the use of antibodies as reagents. Highly specific binding and large association constants (K_a) of antibodies make them useful for the detection and quantitation of analytes in complex biological matrices such as clinical specimens. Methods based on immunochemical reactions are some of the most sensitive and specific assays available in clinical chemistry. Most immunoassays have been automated and are widely used in the clinical laboratory [1,2].

Antibodies belong to a group of glycoproteins called immunoglobulins that demonstrate highly specific binding to analyte molecules, also called *antigens* [3]. The site where the antibody binds to the antigen is called an *epitope*; each antigen may have multiple epitopes depending on its size and molecular structure. Antibodies can be classified as either polyclonal or monoclonal. The term *polyclonal antibodies* refers to a heterogeneous group of antibodies from multiple B-cell clones that are all specific for the same antigen but bind to different epitopes with varying affinity. *Monoclonal antibodies* are derived from a single B-cell clone and are more specific in that they are directed toward not only a certain antigen but also a specific epitope on that antigen.

There are five different types of immunoglobulins characterized by their function and they are abbreviated as IgA, IgD, IgE, IgG, and IgM; IgG is primarily the type of immunoglobulin used as an immunoassay reagent. Immunoglobulin G is a glycoprotein composed of multiple subunits with a molecular weight of approximately 160 kD; it consists of two identical heavy chains and two identical light chains that are joined by disulfide bonds (see Fig. 12.1). The amino acid sequence at the carboxyl terminus is conserved regardless of the antibody specificity and is known as the constant region of the antibody. Conversely, the amino acid sequence at the amino terminus is highly variable, and this is the portion that confers antigenic specificity to the antibody. The classic "Y" shape of the antibody gives it a single constant region with two variable regions that allow one antibody to bind two antigens, barring steric interference.

The structure of an antibody can be further characterized by how it is cleaved with proteolytic enzymes. The enzyme papain cleaves IgG in the "hinge" region of the antibody and divides the antibody into the three sections labeled in Fig. 12.1. The constant region, or "stem" of the Y, is one fragment and is known as the Fc fragment—this has no antigen-binding properties. The remaining two fragments are the "arms" of the Y and are identical antigen–binding fragments known as Fab fragments—these have one binding site each. An alternative enzyme, pepsin, can be used to digest antibodies, which yields a different fragmentation pattern. No Fc fragment is produced in this digestion, because the constant region of IgG is digested completely. The fragment that remains after digestion consists of both binding sites and is known as an $F(ab)_2$ fragment, which is identified in Fig. 12.1.

Antibodies are produced by the immune system from B lymphocytes when a molecule called an *immunogen* is introduced into a foreign host. An immunogen can be a protein, or sometimes an analyte of interest or synthetic peptide coupled to a large carrier molecule. In order for a substance to be immunogenic, it must have three key attributes: (1) it must possess randomness of structure; (2) it must maintain some areas of structural stability; and (3)

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00012-0

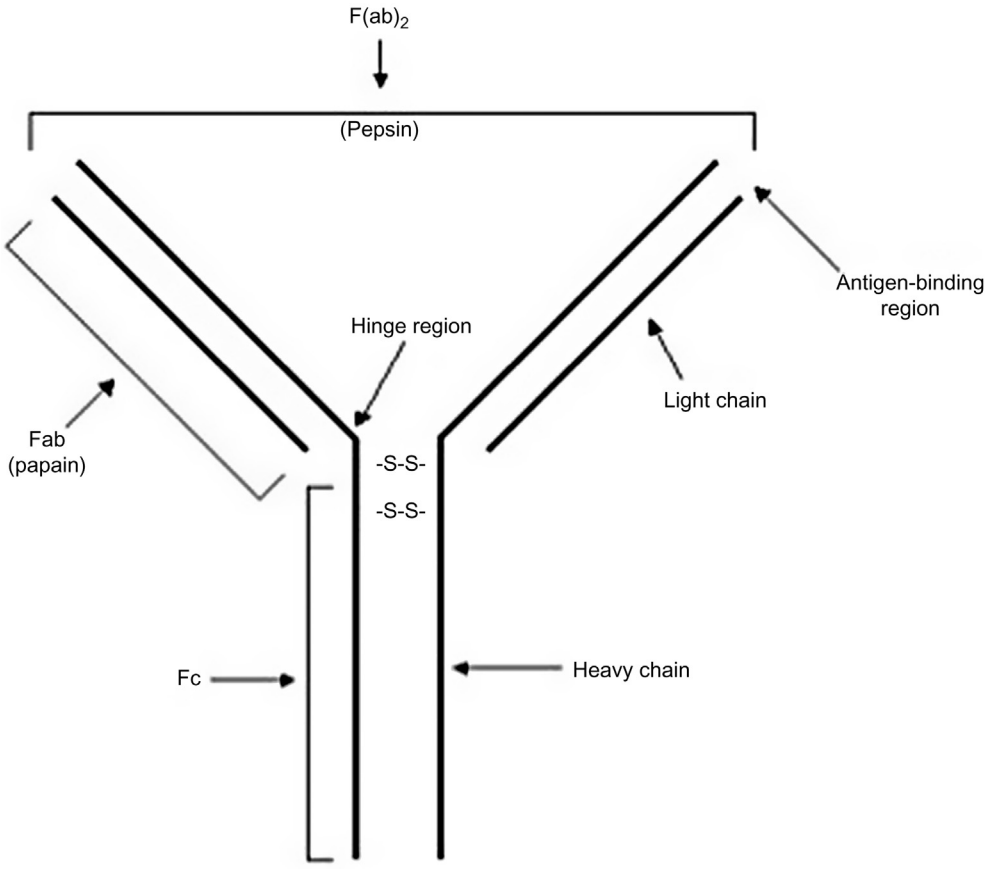


FIGURE 12.1 Basic antibody structure.

it must have a molecular weight of at least 4−5 kD and be recognized as foreign by the host species. An analyte/antigen that is unable to induce an immunogenic response on its own is called a *hapten* and must be coupled with a large carrier protein, such as keyhole limpet hemocyanin or bovine serum albumin, to stimulate a response.

Antiserum is the product of immunogenic stimulation that contains antibodies directed toward the antigen of interest. Polyclonal antiserum is derived by simple introduction of an immunogen to a normal animal host, typically rabbits or goats, which produces a heterogeneous mixture of antibodies from many cell clones directed against a wide variety of epitopes on the antigen of interest with varying affinities and isotypes. Nonspecific antibodies can also be produced. Antisera may be used directly or in a diluted form, or it may undergo purification to isolate only immunoglobulins. Antibodies can be isolated with techniques that yield total IgG or by immunoaffinity using a chromatographic support with immobilized antigen, or immobilized protein A or protein G. Use of immobilized antigen allows the isolation of only IgG that is specific for the analyte of interest.

Monoclonal antiserum is the product of a single clone or plasma cell line, so all the antibodies produced are identical and bind to the same epitope. Production of monoclonal antibodies begins by immunizing a host animal (usually a mouse) just as in polyclonal antibody production. Next, sensitized lymphocytes from the spleen of the immunized animal are fused with an isolated myeloma cell line, which consists of immortalized B-cells. The myeloma cell line is deficient in the enzyme hypoxanthine guanine phosphoribosyl transferase, meaning they cannot synthesize purine bases from thymidine and hypoxanthine in the presence of aminopterin (a synthetic enzyme that blocks tetrahydrofolate synthesis). The spleen cells, however, are not deficient in the enzyme, so when the fusion mixture is cultured in a selective medium containing hypoxanthine, aminopterin, and thymidine, only the fused cells survive. The surviving cell lines are screened for binding to the antigen of interest, with the end product being immortalized cell lines that produce antibodies of a single, known binding specificity.

Antibody−antigen binding is typically described in terms of binding affinity and avidity. *Affinity* refers to the

strength of binding for one site and can be described in terms of the energy of interaction for a single antibody-binding site and its corresponding epitope. *Avidity* refers to the overall binding of one antibody molecule to antigen, describing the sum of binding sites for the whole molecule (e.g., IgG has two binding sites, while an IgM pentamer has 10 binding sites). For example, when comparing an IgG molecule versus Fab fragments obtained from the same clonal population, they would have identical affinities for the antigen of interest, but the IgG would have a higher avidity, because it could bind two antigens and the Fab could only bind one. As a rule, affinity is characterized as a property of the antigen, while avidity is a property of the antibody.

Binding between antibody and antigen is best described as an equilibrium process:

$$\mathrm{Ag} + \mathrm{Ab} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{AgAb}$$

where $k_1 >> > k_{-1}$; depending on the time frame of reference, binding can often be considered irreversible. Antibody−antigen interactions are based on electrostatic (van der Waals) forces, hydrophobic forces, coulombic (ionic) forces, or a combination of the three. The interaction of antibodies and their corresponding antigens is pH- and temperature-dependent, and can be affected by electrolyte concentration as well, depending on which types of interactions are prominent in binding.

Based on an assay on antigen−antibody binding, factors affecting binding can be manipulated to optimize assay sensitivity. Increased temperature will increase the rate of reaction between antibody and antigen, but typically lowers the binding affinity constant—this means that often higher sensitivities can be achieved at lower incubation temperatures due to an increase in binding affinity and a decrease in proteolysis. For protein-bound analytes, a higher temperature may give a higher sensitivity, because it will encourage dissociation of the analyte from the binding protein, while affecting the antibody binding to a lesser degree. Incubation time is also an important parameter to optimize. Longer incubation times allow the reaction to come to equilibrium and let everything that is going to bind to the antibodies have a chance to do so. However, this can lead to long analysis times, as well as decreases in sensitivity when the labeled antigen is sensitive to degradation. To avoid these issues, assays can be performed under nonequilibrium conditions with shortened incubation times. Buffer conditions are also important for assay optimization. In some cases, the antibody−antigen binding is pH-dependent—design of the assay must include an examination of how various pH conditions affect the assay. In addition, pH effects may vary with temperature, so both factors must be examined together to arrive at the optimum assay design. Cationic salts inhibit antibody binding with cationic haptens corresponding to decreasing ionic radius and increasing radius of hydration:

$$\mathrm{Cs}^+ > \mathrm{Rb}^+ > \mathrm{NH}_4^+ > \mathrm{K}^+ > \mathrm{Na}^+ > \mathrm{Li}^+$$

The same effect is seen with anionic haptens and anionic salts:

$$\mathrm{CNS}^- > \mathrm{NO}_3^- > \mathrm{I}^- > \mathrm{Br}^- > \mathrm{Cl}^- > \mathrm{F}^-$$

Immunoassays

Using chemical labels coupled to antibodies or antigens, very sensitive immunoassays can be developed that detect much lower concentrations of a substance than other analytical techniques. These assays can be performed in many different formats with a wide variety of detection methods and can be used to detect both large and small analytes. Immunoassays can be characterized with regard to whether a separation step is necessary, whether the antigen or antibody is labeled, or by the type of label that is used for sample detection. Immunoassays are simply one type of ligand-binding assays, which may use receptors, carrier proteins, and others in addition to antibodies for binding the analyte of interest. For example, intrinsic factor and folate-binding protein are used in competitive protein-binding assays for serum vitamin B12 and folate, respectively.

Homogenous versus heterogeneous immunoassays

Heterogeneous immunoassays are assays that require physical separation of the labeled antibody−antigen complex from the unbound components (antibodies and/or free antigen) for analysis [4]. During the separation step, heterogeneous assays assume that antibody−antigen binding is essentially irreversible during the separation step, since k_1 is much greater than k_{-1} (the amount that dissociates is very small). One way this separation happens is through precipitation of the labeled antigen−antibody complex, either using a precipitating chemical or by cross-linking with other antibodies, and analyzing the precipitate to quantify the amount of bound label. In addition, adsorbents such as activated charcoal or silica particles can be added to solution to bind free antigen for subsequent removal by centrifugation. A common approach is to bind the antibodies to a solid support and then remove the unbound molecules through various wash steps. In an enzyme-linked immunosorbent assay (ELISA), the solid support is the wall of a microwell plate, while in automated immunoassay systems, the solid support is often a magnetic bead or particle. Separation

techniques such as size-exclusion or ion-exchange chromatography, as well as electrophoresis, can also be used for heterogeneous immunoassays.

Homogeneous immunoassays are able to distinguish between the free and bound label for detection without a physical separation step [4]. This is accomplished by selecting a label that is modified by antibody binding, whether it results in an increase or decrease of signal. For this type of assay to be useful, it is important that the change in signal due to antibody binding is proportional to the concentration of analyte in the specimen. This technique allows a combination of the label incubation and detection of analyte into one simple step. The homogenous approach is useful, because it eliminates the wash step and is very compatible with most automated chemistry analyzers.

Competitive immunoassays

Competitive immunoassays are based on immunochemical reactions in which the number of antigen-binding sites is limited. Both the antigen of interest and a labeled analog are incubated with a known amount of antibody, and the signal produced is based on the competition between the analyte and the corresponding label for the antibody-binding sites [5]. The amount of labeled antigen bound to the antibody is inversely proportional to the amount of analyte in the sample and can produce a calibration curve, as seen in Fig. 12.2A.

The most common approaches to competitive immunoassays are the simultaneous or sequential addition techniques. In *simultaneous addition*, both the label and sample (or calibrator) are added to the reaction mixture at the same time. After a fixed period of time, the amount of bound label is determined and used to construct a calibration curve or determine the concentration of the unknown. For optimum results using this technique, the differences in antibody-binding affinity due to matrix effects for both the sample antigen and the labeled antigen should be minimal. In addition, it is important that the probability of binding for both the labeled antigen and the sample antigen is similar. *Sequential addition* is a valuable technique when a lower limit of detection is needed. First, the sample is incubated with the antibody until equilibrium is reached; then, label is added and comes to equilibrium, followed by the detection of bound label. Importantly, the length of time needed for incubation will be determined by the type of interactions between the Ab−Ag and the relative amounts of each. The sequential approach to competitive immunoassays can lower the limit of detection up to fivefold and is also useful if there are slight differences in affinity for the analyte antigen and the labeled analog.

An alternative approach to a competitive immunoassay is the displacement immunoassay. In this technique, reagent antibody is saturated with labeled analog followed by addition of sample. The analyte in the sample displaces the label, and the displaced (free) labeled compound is measured rather than the bound label. This type of assay is dependent on local equilibria, so if $k_1 >> k_{-1}$, then less label is displaced in contrast to the situation where $k_1 \geq k_{-1}$. The signal for displacement immunoassays is directly proportional to the analyte concentration (Fig. 12.2B), in contrast to the inversely proportional signals seen in the simultaneous and sequential addition immunoassays.

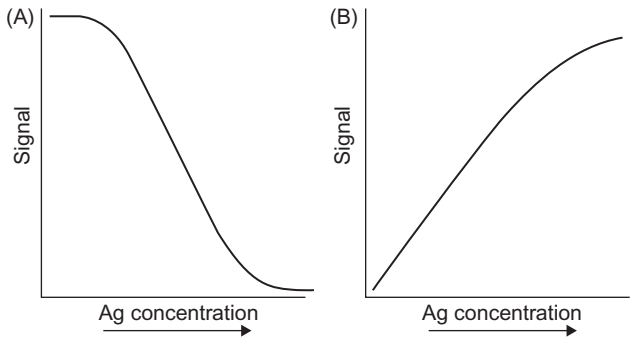


FIGURE 12.2 Immunoassay calibration curves: (A) competitive binding assay and (B) noncompetitive binding assay.

Noncompetitive (immunometric) immunoassays

Noncompetitive immunoassays, or immunometric assays, are based on immunochemical reactions, where there is reagent excess [5]. In noncompetitive immunoassays, the signal is directly proportional to the amount of analyte in the sample, producing a similar curve to that obtained with a displacement immunoassay (Fig. 12.2B). The most common noncompetitive immunoassay format is the *sandwich immunoassay*, where two antibodies are used for analyte detection. Both polyclonal and monoclonal antibodies are suitable for use in sandwich immunoassays. However, in order to use this format, the antigen of interest must be large enough to bind simultaneously two separate antibodies. The first antibody in the reaction is called the *capture antibody*, which is either adsorbed or covalently linked to a solid support and is used to extract the analyte from the sample matrix. A second labeled antibody, known as the *tracer* or *conjugate antibody*, reacts with the captured antigen at a separate and distinct epitope from the capture antibody and is used to detect the analyte when all other components have been washed away. This type of assay can be performed either in a sequential or simultaneous incubation format, although for simultaneous incubation two distinct monoclonal antibodies for different epitopes must be used to prevent competition.

Immunoassay detection methods

There are many different types of labels, for both antigen and antibodies, that can be used in immunoassays. Choice of label is dependent on the analytical sensitivity desired for the measurement, as well as the immunoassay format. A list of label types along with typical detection limits associated with them can be seen in Table 12.1. The earliest immunoassays used isotopic labels such as ^{125}I, ^{131}I, ^{3}H, ^{57}Co, and ^{32}P for antigens in competitive immunoassays and for antibodies in sandwich immunoassays—these were appropriately termed radioimmunoassay (RIA) and immunoradiometric assay. The appeal of radioactive labels is based on the low limits of detection achieved due to very low background interference. Radioactive labels were also used to develop an early homogeneous assay [6] based on the modulation of ^{3}H or ^{125}I labels by scintillant-laden microparticles. However, RIAs are now used infrequently in clinical laboratories due to concerns regarding safe handling and disposal of radioactive materials, as well as the inherent instability of radiolabels.

Currently, nonisotopic labels are more widely used in immunoassays because of the drawbacks that come with handling radioactive labels. As early as the 1970s, enzyme labels for immunoassays were developed, and since then, a wide variety of nonisotopic labels have been established. These types of labels include enzymes, fluorescent molecules, chemiluminescent labels, and even electrochemical labels as well as many other newly developed, unique labels. Some of these labels are able to achieve detection limits well below those possible with radiolabels, and none of them pose the problems of handling and disposal that occur when working with radioactive materials. Increased specificity can be gained in an immunoassay by coupling a biotin–avidin interaction with the antibody–antigen interaction by using biotin-conjugated antibody or antigen and then introducing labeled avidin to the reaction mixture. In some applications, multiple labels are utilized by taking advantage of the multiplicity of avidin–biotin binding. One avidin molecule can bind four biotin molecules, so this complex scheme would involve a biotin conjugate followed by addition of avidin, which could then bind three more labeled biotin molecules. It should be noted that in assays utilizing biotin–streptavidin interactions, biotin supplementation may cause falsely elevated or decreased results depending on assay design.

TABLE 12.1 Immunoassay label detection limits.

Label type	Detection limit (M)
Fluorescence	1×10^{-10}
Chemiluminescence	1×10^{-13}
Fluorescence polarization	1×10^{-14}
Radioactive	1×10^{-15}
Time-resolved fluorescence	1×10^{-17}
Electrochemiluminescence	2×10^{-20}
Enzyme-based	
Photometric	1×10^{-16}
Fluorescence	1×10^{-19}
Coupled enzyme reaction	1×10^{-20}
Chemiluminescence	1×10^{-21}

Fluorescence

Fluorescent labels are commonly used for competitive binding and sandwich immunoassays in much the same manner as radioisotopes. The most commonly used label is fluorescein isothiocyanate (FITC), but dyes such as Lucifer Yellow or Rhodamine B can also be employed as fluorescent labels. Detection limits using fluorescence range from 0.1 pM to 1 nM, which is sensitive enough to use for low concentration analytes, but avoids the handling and disposal issues that come with using radioactivity. However, a major challenge with fluorescence detection is that many substances found in patient samples (e.g., drugs and drug metabolites, proteins, and bilirubin) exhibit fluorescence and can cause a high background and interference. One way to avoid this interference is the use of *time-resolved fluorescence* with labels derived from lanthanide elements (i.e., europium, terbium, and samarium). The fluorescence lifetime for these lanthanide chelates is much longer than typical fluorescent molecules, so after excitation, a delay time (400–800 ms) is introduced to let background fluorescence decay, and then the signal may be measured [4].

Fluorescence detection can be used in homogenous immunoassays by employing a fluorescence polarization immunoassay (FPIA) format. This approach is based on competitive binding of antigen in the sample with a fluorescein-labeled antigen reagent [4]. An FPIA uses the changes in polarization of fluorescence due to molecular rotation in solution to determine the extent of label–antibody binding. When the labeled antigen is unbound, it rotates very rapidly and therefore has a low degree of polarization; but when the label is bound to an antibody molecule that rotates at a much lower rate, the degree of polarization is increased (see Fig. 12.3). The change in polarization is proportional to the concentration of antigen in the sample, and because the label behavior is modified by antibody binding, no separation step is needed. FPIA assays were once widely used for drug measurement, primarily in the Abbott TDx analyzer, but their use in

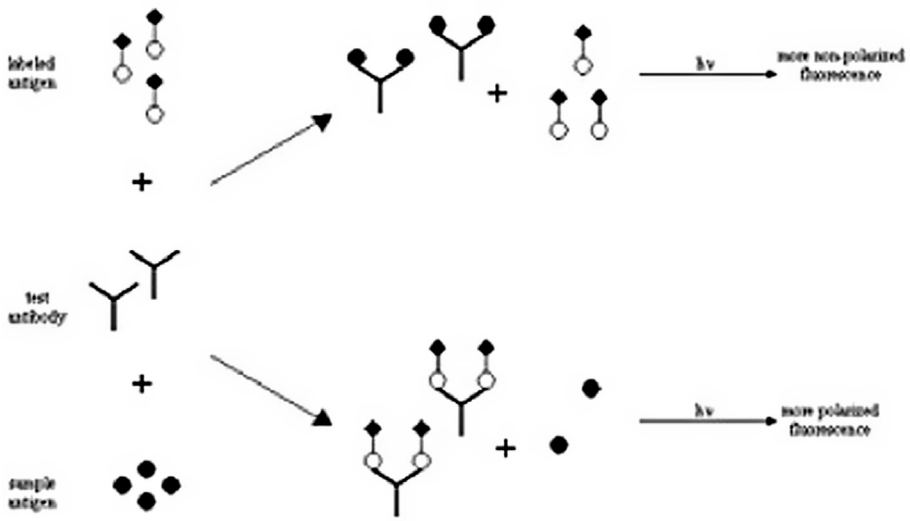


FIGURE 12.3 Fluorescence polarization immunoassay.

clinical chemistry has significantly decreased since that analyzer was removed from the market in 2011.

Enzymes

Enzyme-based labels such as alkaline phosphatase, horseradish peroxidase, glucose-6-phosphate dehydrogenase, and β-galactosidase are frequently used in quantitative immunoassays [4,7]. One significant advantage for enzyme labels is the amplification of signal; one enzyme label can produce many reporter molecules, up to 100-fold amplification (using a fluorescent substrate can produce amplification up to 100,000-fold). The substrate for enzyme detection systems can be chosen to produce a colored product for photometric monitoring, a fluorescent product such as umbelliferone, or even to produce an intermediate substrate for a second enzymatic reaction to produce further amplification. The most common heterogeneous enzyme immunoassay is the ELISA, which is an immunometric assay where the captured antibody is bound to a solid support and the tracer antibody is enzyme-labeled [4]. After the capture:antigen:tracer complex is formed, all other sample components are washed away and substrate is added; the amount of product formed is proportional to the concentration of antigen in the sample (Fig. 12.4). An ELISA can also be used to detect antibodies of a given specificity by using antigen linked to the support, adding sample and using labeled antibodies to IgG as the tracer (e.g., detection of anti-HCV antibodies).

There are two commonly used types of homogeneous enzyme immunoassays: the *enzyme-multiplied immunoassay technique* (EMIT) and the *cloned enzyme donor immunoassay* (CEDIA) [4,8]. In an EMIT assay, the patient specimen or calibrator is incubated with an enzyme-labeled antigen and a fixed amount of antibody specific for the antigen of interest as well as enzyme substrate (Fig. 12.5) [4,8]. Antigen in the sample and the labeled antigen compete for antibody-binding sites; when antibody is bound to the labeled antigen, the enzyme is inhibited, because it is blocked from interaction with the substrate. The change in enzymatic activity is proportional to the concentration of antigen in the patient sample or calibrator (e.g., as the concentration increases, so does the enzyme activity).

A CEDIA assay is also a competitive immunoassay [8,9], but it is more complex than the EMIT assay. Using genetic engineering techniques, two inactive fragments of β-galactosidase are produced by manipulating the Z gene of the *lac* operon of *Escherichia coli*. These two fragments are able to reassemble to form an active enzyme, even if one of the fragments is covalently linked to another molecule. The competitive antigen in the assay is labeled with one fragment of the enzyme and is incubated with patient sample, along with the second enzyme fragment, the enzyme substrate, and a fixed amount of antibody directed toward the antigen of interest (Fig. 12.6). Antibody that binds to labeled antigen prevents reassembly of the enzyme and no reporter molecule is produced; as the sample antigen concentration is increased, the enzyme activity is also increased. As in the EMIT assay, the change in enzymatic activity is directly proportional to the concentration of antigen in the sample.

Chemiluminescence

Chemiluminescent labels are based on the emission of light produced during a chemical reaction. These labels are very useful, because they provide very low limits of detection with little to no background interference; however, in these assay formats, the reaction conditions must be carefully controlled. Chemiluminescent labels exist as

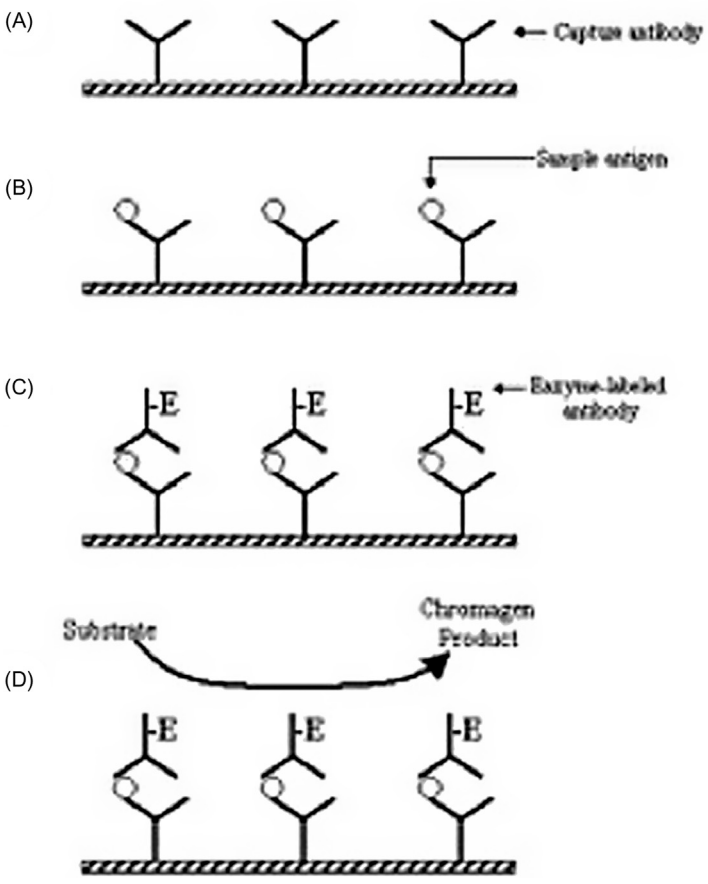


FIGURE 12.4 Enzyme-linked immunosorbent assay.

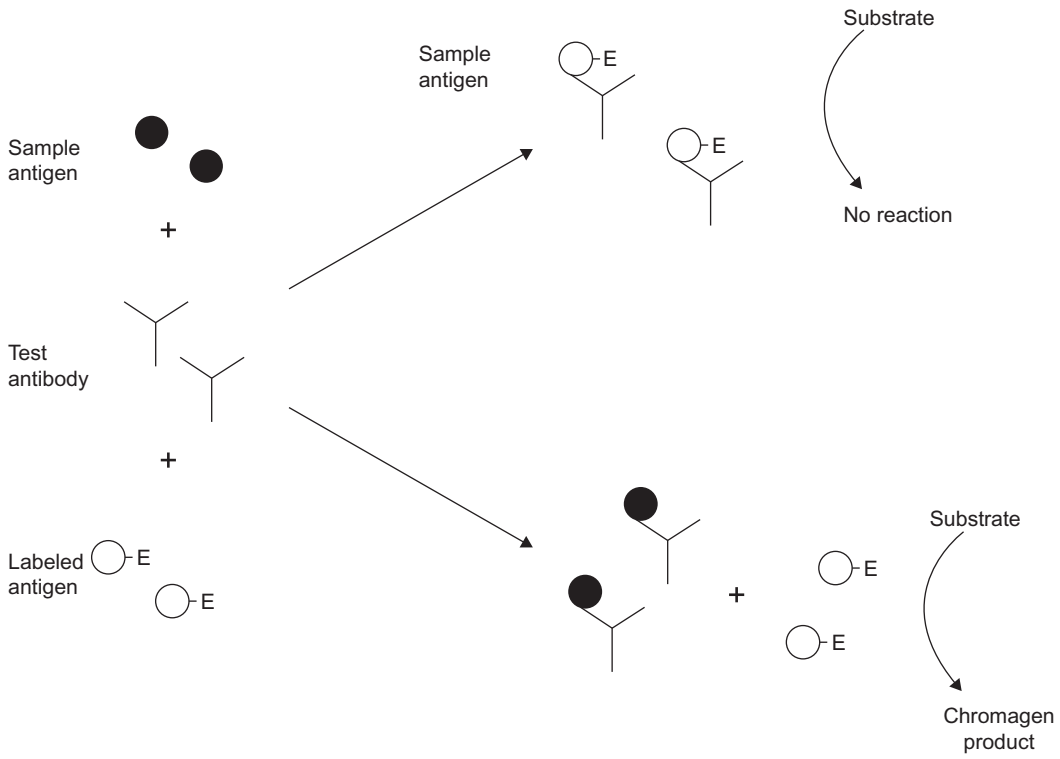


FIGURE 12.5 Enzyme-multiplied immunoassay technique.

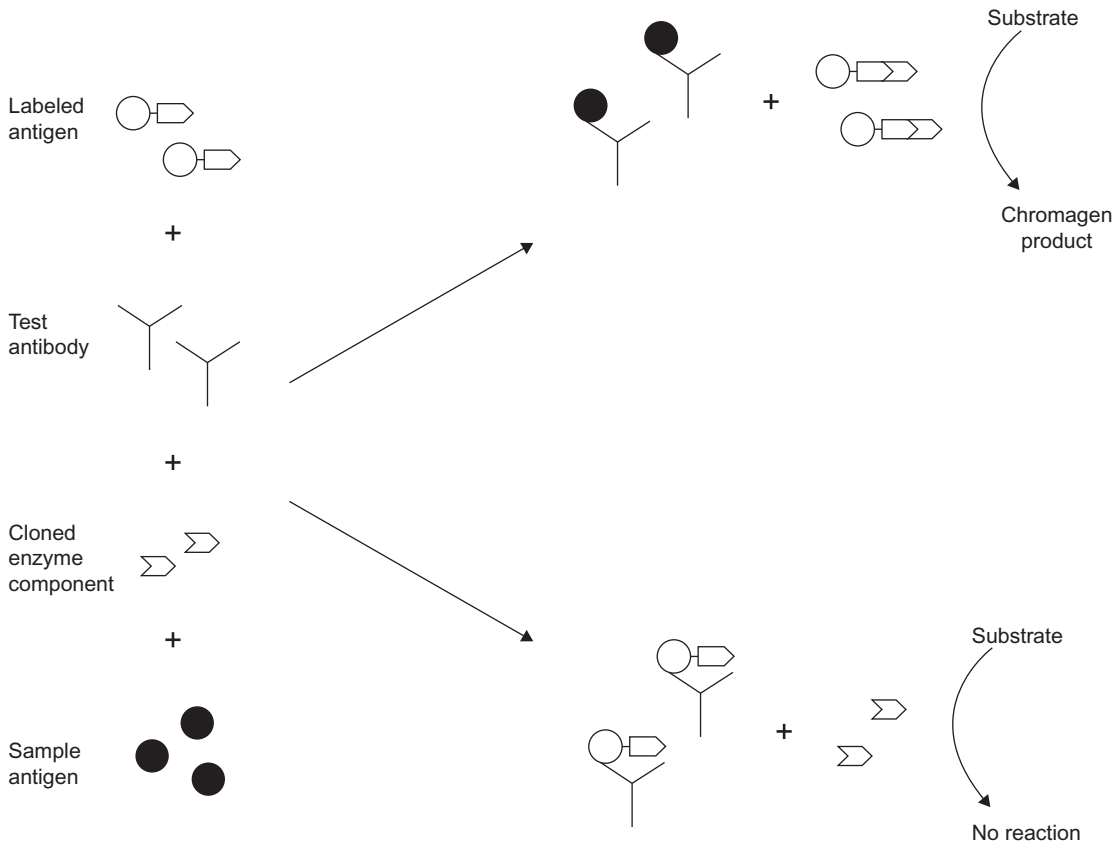


FIGURE 12.6 Cloned enzyme donor immunoassay.

both biologically derived and chemically manufactured compounds [10]. Biologically derived chemiluminescence labels include firefly luciferase and apoaequorin, a protein derived from the bioluminescent jellyfish *Aequorea victoria*. Firefly luciferase activates D-luciferin substrate by oxidation in the presence of ATP, molecular oxygen, and magnesium ions to emit light at 560 nm, while apoaequorin is activated by reaction with coelenterazin and calcium ions to emit light at 469 nm.

Nonbiologic compounds can also be used as chemiluminescent labels, including isoluminol and acridinium esters [11]. Isoluminol is oxidized in the presence of a catalyst such as microperoxidase, producing a relatively long emission at 425 nm. Acridinium esters are highly sensitive labels that are activated by oxidation using hydrogen peroxide in alkaline solution to give a brief flash of light at 429 nm. Chemiluminescence can be made even more sensitive when combined with other detection methods. For instance, electrochemiluminescence utilizes a molecule such as ruthenium (II) tris(bipyridyl) that undergoes chemiluminescence at a given redox potential, producing mass detection limits as low as 20×10^{-21} moles [10,12]. In addition, chemiluminescent substrates can be combined with enzymatic labels to take advantage of enzyme amplification effects and the inherent sensitivity of chemiluminescence by using a substrate known as disodium 3-(4-methylspiro-[1,2-dioxetane-3,2′-tricyclo-[3.3.1.1]decan]-4-yl) phenyl phosphate to obtain mass detection limits as low as 1×10^{-21} moles [13].

Particle-based immunoassays

Some automated immunoassays are based on antibodies or antigens linked to solid, spherical microparticles [14]. In some cases, such as the homogeneous particle-enhanced turbidimetric inhibition immunoassay, the microparticles are linked to the analyte of interest, and when antibodies are added to the reagent solution at an optimized ratio, the microparticles agglutinate and form insoluble complexes that scatter light. Competing analyte from a patient specimen (or control specimen) inhibits the formation of the insoluble complexes at a rate proportional to the concentration of the analyte. A similar mechanism is responsible for analyte quantitation in the kinetic interaction of microparticles in a solution immunoassay platform. An example of a microparticle-based light scattering immunoassay can be seen in Fig. 12.7. Measurement of light scattering immunoassays by turbidimetry and nephelometry is described in Chapter 7, Spectrophotometry.

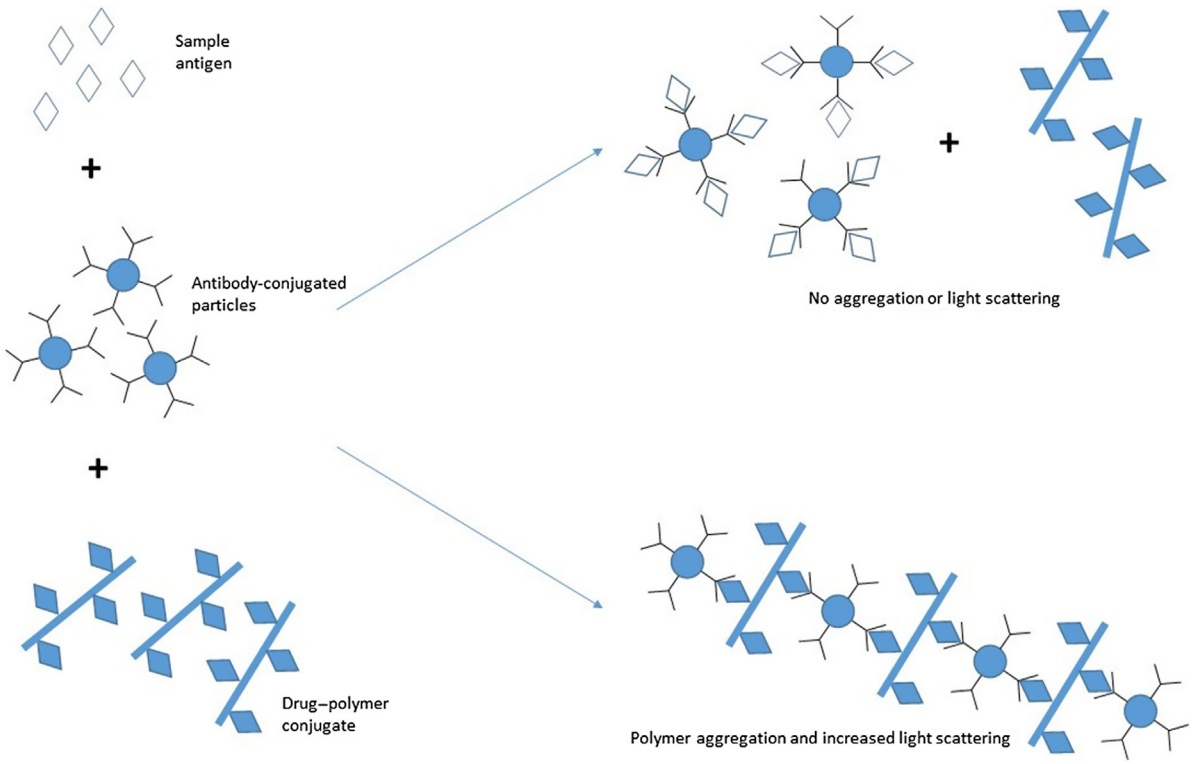


FIGURE 12.7 Particle-based light scattering immunoassay.

An alternative use of microparticles for heterogenous assays is microparticle enzyme immunoassay. In an immunometric format example, a sample is incubated with antibody- or antigen-coated microparticles and enzyme-labeled conjugate. An aliquot is filtered through a matrix such as a glass-fiber capture surface. Conjugate can alternatively be added at this step following a wash. The surface is washed and enzyme is added with the signal generated in proportion to the analyte concentration.

Alternative labels and technologies

Alternative detection techniques for nonisotopic immunoassays have been developed in response to demands for lower detection limits, an increase in label specificity, or increased flexibility in applications. One example is the *enzyme channeling immunoassay* that combines the simplicity of a homogeneous immunoassay with the specificity of a sandwich immunoassay [15]. In this approach, the antigen of interest binds two antibodies that are labeled with different enzymes; the enzyme reactions are coupled and signal is produced when the enzymes are in close proximity due to binding within the immune complex. A similar technique is the *fluorescence resonance transfer immunoassay* (FRET), which utilizes coupled fluorescence reactions [16,17]. In the reagent mixture of competitive assays, the labeled antigen contains a fluorophor donor, while the antibody is labeled with a fluorescent acceptor; when the labeled antigen is antibody-bound, a fluorescent signal occurs; as the analyte concentration increases, the signal decreases. Time-resolved FRET forms the basis for the time-resolved amplified cryptate emission methodology on the Kryptor instrument (Thermo Fisher Scientific) that is currently used for procalcitonin measurement. Another related homogeneous immunoassay technique is luminescent oxygen-channeling immunoassay, a sensitive methodology employed on the Siemens Dimension Vista. In this immunometric assay format, antibody-coated sensitizer and chemiluminescent beads form a sandwich with the antigen. When the photosensitizer containing bead is exposed to light, the singlet oxygen produced diffuses into the associated chemiluminescent compound-containing particle, which initiates a chemiluminescent reaction [18].

Many other methods have been developed, including those that incorporate polymerase chain reaction amplification with DNA or RNA labels, optical diffraction detection, and surface physiochemical detection techniques for biosensors as well as those with single molecule detection and ultrasensitive properties [19–22].

Immunoassay interferences

Although there are many advantages inherent to immunoassays, it is important to consider some of the factors

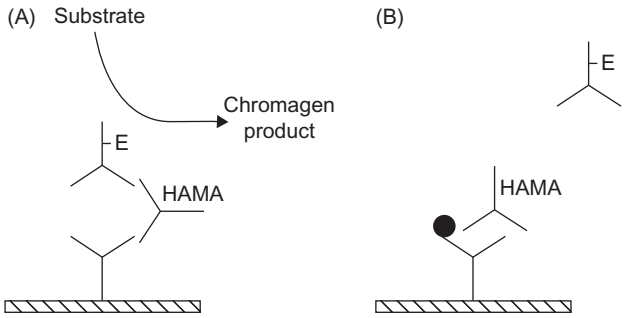


FIGURE 12.8 Heterophilic antibody interference: (A) positive interference and (B) negative interference.

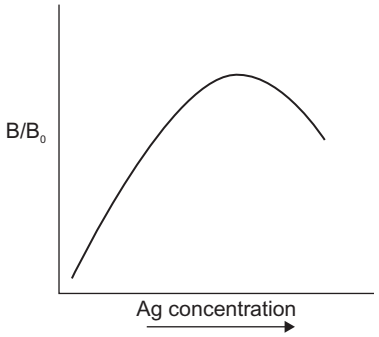


FIGURE 12.9 Calibration curve demonstrating the hook effect.

that may interfere with an immunoassay. These include heterophilic antibodies from the patient, assay artifacts caused by large analyte concentrations (hook effect), and a lack of standardization between immunoassays for the same analyte from different manufacturers. Each of these interferences can influence both immunoassay performance and utility.

In sandwich immunoassays, one particular problem is interference from circulating heterophilic human antibodies specific for animal immunoglobulins when they are used as reagents. The most common type of these is human antimouse antibodies (HAMA) that are often present in the blood of patients that have been given mouse monoclonal antibody imaging or therapeutic agents, or have been exposed to other mouse antigens [23]. Heterophilic antibodies can cause false-positive results by cross-linking the capture and label antibodies in a manner similar to the analyte of interest or they can cause false negative results by blocking analyte binding to the capture or label antibody (Fig. 12.8). The presence of HAMA or other heterophile antibodies can be confirmed through dilution experiments that do not give linearly proportional results, or retesting of a sample after incubation with animal serum (e.g., mouse antisera for HAMA) or other blocking agents. Nonimmune animal serum is often included in immunoassay kits using animal antibodies as reagents to minimize the effects of heterophilic antibodies. In addition, many antibody-based therapeutic agents are now derived from antibody fragments or chimeric antibodies (i.e., fusion proteins genetically engineered to make animal produced proteins mimic human proteins) in order to avoid heterophilic antibody reactions.

Another problem that can arise and give falsely low results for immunometric assays is known as the "hook effect" [24]. This phenomenon occurs when the concentration of analyte is so high that both the capture and label antibodies are saturated. As a result, the assay response will decrease at high antigen concentrations—the response will drop off after a maximum response giving the curve a characteristic "hook" shape (see Fig. 12.9). This effect leads to falsely low patient results when the concentration of an analyte is elevated far above the analytical range of the assay and can be discovered by dilution followed by sample retesting. This type of error can be avoided by using a large excess of capture and label antibodies with respect to the analytical range of the assay.

A different sort of problem arises when immunoassays are developed for the same analyte by different researchers or manufacturers. Each developer uses a different antibody directed to a unique epitope that may exhibit varying binding properties, leading to different responses for the same concentration of analyte. In an isolated environment, this is not a problem as long as the institution uses only one type of assay; however, the establishment of a standard reference interval for the analyte becomes much more difficult, because various institutions often use different assays for the same analyte. A good example of this is demonstrated in the development of a cardiac troponin I (cTnI) assay. cTnI is present in the circulation in three forms: free cTnI, a two-unit complex with cardiac troponin C (cTnI-cTnC), and a three-unit complex with cTnC and cardiac troponin T. Various assays fail to agree with each other, because the reagent antibodies for each assay recognize different epitopes.

Therefore common antibodies in addition to a primary reference material are needed for standardization of the method [25,26].

Advances in immunoassay testing

Although immunoassays are some of the most widely used tests in the clinical laboratory today, they do suffer from multiple drawbacks and are subject to interferences. In recent years, a number of advancements have led to progress in the development of new methods utilizing antibodies, which have been able to circumvent some of these limitations. Three such examples are described below.

HIV Ag/Ab combination assay

For HIV diagnosis, the seroconversion window corresponds to the period between HIV infection and the

presence of detectable antibodies to the virus. This period is 2−4 weeks after infection. During this initial phase, the virus is actively replicating, produces high titers in the blood, and can produce an acute HIV syndrome. This initial phase is a period of high transmission, and acutely infected individuals, who may be unaware of their status, are responsible for a substantial fraction of newly transmitted infections. Despite the multiplicity of changes taking place in this early phase, anti-HIV titers remain low in the blood during this period. Thus traditional tests that only detect HIV antibodies are simply not adequate to identify and diagnose these individuals.

The HIV Ag/Ab combination assay is a fourth-generation assay that allows for the simultaneous detection of p24 antigen and anti-HIV antibodies. Abbott was first to receive FDA approval for this test in the United States in 2010. Since then, other companies have also commercialized products, including Siemens, BioRad, and Alere. These fourth-generation assays are similar in that they can all detect both acute and long-term infection. The former is manifested through p24, a component of HIV that is detectable in the blood during the acute phase. Abbott's assay is a chemiluminescent microparticle immunoassay. The assay provides a signal/cutoff value, which reaches detectable levels in the presence of either p24 or anti-HIV antibodies. This assay will not allow for the discrimination in signal origin. Alere's Determine HIV-1/2 Antigen/Antibody Combo is a lateral flow, point-of-care immunoassay and, in contrast to the Abbott test, is not performed on an automated analyzer. It, however, contains distinct zones on the test strip, which allow for the distinction between the signal generated from p24 and that from the anti-HIV antibodies. Most recently, the BioRad Bioplex 2200 also has received FDA approval for an automated assay to detect HIV p24 antigen and anti-HIV antibodies and can not only detect these analytes but can also distinguish between them.

For additional details, see Chavez et al. [27] and references therein.

Mass spectrometry measurement of thyroglobulin

For patients with differentiated thyroid cancers, the use of serum thyroglobulin testing has been the mainstay of surveillance regimens to identify residual disease and to detect recurrence. Historically, immunoassays have been the method of choice for this analysis; however, there are some significant challenges encountered with current Tg immunoassays. For example, the presence of anti-Tg autoantibodies in 10%−30% of patients can result in falsely elevated or depressed levels of this marker in routine testing. One approach to circumvent such problems with this immunoassay is by adding a recovery step. The addition of a defined amount of Tg to the sample and then calculating its recovery after measurement enables the laboratory to determine whether anti-Tg antibodies are present in the sample. This additional step provides an opportunity to limit the reporting of inaccurate results. However, such recovery experiments are not recommended in the most recent guidelines of the American Thyroid Association [28].

Alternately, mass spectrometry can be used to circumvent some of the interference issues with Tg immunoassays. Hoofnagle et al. [29] and Hoofnagle and Roth [30] have developed such an approach. In this published method, the authors utilize a peptide immunoaffinity enrichment followed by liquid chromatography−tandem mass spectrometry (LC-MS/MS) analysis; using this approach, they have successfully demonstrated the analytically sensitive detection of single peptides of thyroglobulin in human serum samples. Their procedure includes a step whereby all proteins within the sample are digested into peptides. In contrast to the immunoassay approach that detects epitopes on the full length protein and is susceptible to being masked by an interfering antibody, their approach captures and detects peptides and is not subject to interferences at the protein level. Further developments in their assays and subsequent prospective evaluation will determine whether these technical advancements can directly translate into improved patient outcomes.

Multiplexed immunoassay testing

Relative to traditional single-marker assays, multianalyte panels offer the advantage of testing multiple analytes at once, allowing for the reduction in cost, time spent on testing, and specimen volume requirements. They have been developed for a number of applications including autoimmune testing via antinuclear antibody detection/characterization [31], autoantibody profiling for antiphospholipid syndrome [32], HIV testing, and other diseases. In each of these cases, the presence/absence or quantitation of multiple analytes is useful for clinical decision-making and patient management. Several successful FDA approved diagnostics have become available over the last few years, including those from manufacturers such as BioRad, Biomerieux, BioSite, Focus Diagnostics, and Randox.

The advantages conferred by these assays may translate into improved test performance characteristics such as higher sensitivity and/or specificity. In order to make this testing practical in a clinical laboratory setting, a multiplex testing system must be able to process samples in high throughput, support random access testing of specimens, and lend itself to automation. The BioRad Bioplex 2200 analyzer is one such platform—it is a fully

automated, random access, multiplexed immunoassay analyzer capable of providing both quantitative and qualitative assessments of antibody measurements. Its testing methodology is based on the Luminex X-MAP technology for multiplex analyte profiling. Briefly, it utilizes fluoromagnetic beads as the solid phase and these beads are coated with antibodies directed to various analytes. The different beads fluoresce as spectral variants and allow for the simultaneous measurement of many different antigens simultaneously in a single sample. It is one example that highlights the advantages of using multiplex assays on a single platform suitable for clinical testing and a powerful tool in the armamentarium of the clinical laboratory.

Point-of-care immunoassays

The most widely used point-of-care testing (POCT) immunoassay is used for the detection of human chorionic gonadotropin (hCG), more commonly known as a "home pregnancy test" or qualitative pregnancy kit. Many different manufacturers produce these tests, but all the tests operate on the same basic principle. A urine specimen is applied to a porous membrane that draws the liquid from the application point to immobilized anti-hCG positioned further up the membrane. If hCG is present in the sample, it will be bound to the immobilized antibodies, followed by a detection step. Detection steps can be varied, depending on the manufacturer. One approach is to add a second labeled antibody solution that will bind to the captured hCG, followed by a color development step [33]. Another approach is to use colored microparticles for detection (Abbott). In this approach, the initial capture antibody is immobilized to colored beads, which are located at the point of application. As the sample and microbeads migrate across the membrane, they encounter a second capture antibody that will serve as a "detection zone"; if there is hCG present in a sample, a sandwich complex will be formed and a visible positive signal will occur.

Other applications of point-of-care assays are found in the hospital emergency department (ED) and in locations where testing for drugs of abuse is necessary. In addition, point-of-care tests for troponin I are available for use in the triage of patients that present to the ED with acute coronary syndrome [34]. POCT devices based on immunoassays are also being developed to screen for drugs of abuse in oral fluids [35]. In order to be useful, each POCT device must not only produce a clear signal that makes it easy to distinguish between positive and negative but also contain a control mechanism to ensure that the device is working properly (e.g., the specimen moves up the membrane, all reagents are included in the reaction, detect the presence of interfering heterophilic antibodies, etc.).

References

[1] D.W. Chan (Ed.), Immunoassay Automation: A Practical Guide, Academic Press, San Diego, CA, 1992.

[2] D.W. Chan (Ed.), Immunoassay Automation: An Updated Guide to Systems, Academic Press, San Diego, CA, 1996.

[3] E. Liddell, Antibodies, in: D. Wild (Ed.), The Immunoassay Handbook, fourth ed., Elsevier, Oxford, 2013, pp. 345–365.

[4] D.W. Chan (Ed.), Immunoassay: A Practical Guide, Academic Press, San Diego, CA, 1987.

[5] R. Ekins, Toward immunoassays of greater sensitivity, specificity and speed: an overview, in: A. Albertini, R. Ekins (Eds.), Monoclonal Antibodies and Developments in Immunoassay, Elsevier/North-Holland Biomedical Press, New York, NY, 1981, pp. 3–21.

[6] M. Picardo, K.T. Hughes, Scintillation proximity assays, in: J.P. Devlin (Ed.), High Throughput Screening, Marcel Dekker, New York, 1997, pp. 307–316.

[7] T. Porstmann, S.T. Kiessig, Enzyme immunoassay techniques: an overview, J. Immunol. Methods. 150 (1992) 5–21.

[8] S.H. Jenkins, Homogeneous enzyme immunoassay, J. Immunol. Methods. 150 (1992) 91–97.

[9] W.D. Engel, P.L. Khanna, CEDIA® in vitro diagnostics with a novel homogeneous immunoassay technique: current status and future prospects, J. Immunol. Methods. 150 (1992) 99–102.

[10] I. Weeks, L.J. Kricka, D. Wild, Signal generation and detection systems (excluding homogeneous immunoassays), in: D. Wild (Ed.), The Immunoassay Handbook, fourth ed., Elsevier, Oxford, 2013, pp. 267–285.

[11] L.J. Kricka, Chemiluminescent and bioluminescent techniques, Clin. Chem. 37 (9) (1991) 1472–1481.

[12] C. Grimshaw, C. Gleason, E. Chojnicki, J. Young, Development of an equilibrium immunoassay using electrochemiluminescent detection for a novel recombinant protein product and its application to pre-clinical product development, J. Pharm. Biomed. Anal. 16 (1997). 605-2.

[13] I. Bronstein, B. Edwards, J.C. Voyta, 1,2-dioxetanes: novel chemiluminescent enzyme substrates. Applications to immunoassays, J. Biolumin. Chemilumin. 4 (1) (1989) 99–111.

[14] I.A. Dinis-Olivera, Heterogeneous homogeneous immunoassays for drug analysis, Bioanalysis. 6 (21) (2014) 2877–2896.

[15] D.J. Litman, T.M. Hanlon, E.F. Ullman, Enzyme channeling immunoassay: a new homogeneous enzyme immunoassay technique, Anal. Biochem. 106 (1) (1980) 223–229.

[16] E.F. Ullman, M. Schwarzberg, K.E. Rubenstein, Fluorescent excitation transfer immunoassay. A general method for determination of antigens, J. Biol. Chem. 251 (14) (1976) 4172–4178.

[17] E.F. Ullman, Homogeneous immunoassays, in: D. Wild (Ed.), The Immunoassay Handbook, fourth ed., Elsevier, Oxford, 2013, pp. 67–87.

[18] E.F. Ullman, H. Kirakossian, A.C. Switchenko, J.M. Ericson, C.A. Wartchow, M. Pirio, et al., Luminescent oxygen channeling assay (LOCI): sensitive, broadly applicable homogeneous immunoassay method, Clin. Chem. 42 (1996) 1518–1526.

[19] T. Sano, C.L. Smith, C.R. Canto, Immuno-PCR: very sensitive antigen detection by means of specific antibody-DNA conjugates, Science. 258 (5079) (1992) 120–122.
[20] Y.G. Tsay, C.I. Lin, J. Lee, E.K. Gustafson, R. Appelqvist, P. Magginetti, et al., Optical biosensor assay (OBA), Clin. Chem. 37 (9) (1991) 1502–1505.
[21] J.W. Attridge, P.B. Daniels, J.K. Deacon, G.A. Robinson, G.P. Davidson, Sensitivity enhancement of optical immunosensors by the use of a surface plasmon resonance fluoroimmunoassay, Biosens. Bioelectron. 6 (3) (1991) 201–214.
[22] S.K. Fischer, A. Joyce, M. Spengler, T.-Y. Yang, Y. Zhuang, M. Scheel Fjording, et al., Emerging technologies to increase ligand binding assay sensitivity, AAPS J. 17 (1) (2015) 93–101.
[23] G.G. Klee, Human anti-mouse antibodies, Arch. Pathol. Lab. Med. 124 (2000) 921–923.
[24] J.T. Wu, S.E. Christensen, Effect of different test designs of immunoassays on "Hook Effect" of CA 19-9 measurement, J. Clin. Lab. Anal. 5 (1991) 228–232.
[25] F.S. Apple, Standardization of cardiac troponin I assays will not occur in my lifetime, Clin. Chem. (2012) 169–171.
[26] R.H. Christenson, D.M. Bunk, H. Schimmel, J.R. Tate, Put simply, standardization of cardiac troponin I is complicated, Clin. Chem. 58 (2012) 165–168.
[27] P. Chavez, L. Wesolowski, P. Patel, K. Delaney, S.M. Owen, Evaluation of the performance of the Abbott ARCHITECT HIV Ag/Ab combo assay, J. Clin. Virol. 52 (Suppl. 1) (2011) S51–S55.
[28] G.L. Francis, S.G. Wageuspack, A.J. Bauer, et al., Management guidelines for children with thyroid nodules and differentiated thyroid cancer, Thyroid. 25 (7) (2015) 716–759.
[29] A.N. Hoofnagle, J.O. Becker, M.H. Wener, J.W. Heinecke, Quantification of thyroglobulin, a low-abundance serum protein, by immunoaffinity peptide enrichment and tandem mass spectrometry, Clin. Chem. 54 (11) (2008) 1796–1804.
[30] A.N. Hoofnagle, M.Y. Roth, Clinical review: improving the measurement of serum thyroglobulin with mass spectrometry, J. Clin. Endocrinol. Metab. 98 (2013) 1343–1352.
[31] K. Op De Beéck, P. Vermeersch, P. Verschueren, R. Westhovens, G. Mariën, D. Blockmans, et al., Antinuclear antibody detection by automated multiplex immunoassay in untreated patients at the time of diagnosis, Autoimmun. Rev. 12 (2) (2012) 137–143.
[32] R. Tozzoli, D. Villalta, Autoantibody profiling of patients with antiphospholipid syndrome using an automated multiplexed immunoassay system, Autoimmun. Rev. 13 (1) (2014) 59–63.
[33] C.P. Price, D.J. Newman (Eds.), Principles and Practice of Immunoassay, second ed., Stockton Press, New York, 1997.
[34] F. Pagani, C. Serena, C. Bosio, C. Cuccia, M. Panteghini, Evaluation of a rapid bedside immunochromatographic assay for detection of cardiac troponin I in whole blood, Clin. Chem. Lab. Med. 39 (5) (2007) 458–459.
[35] C. Barrett, C. Good, C. Moore, Comparison of point-of-collection screening of drugs of abuse in oral fluid with a laboratory-based urine screen, Forensic Sci. Int. 122 (2001) 163–166.

Self-assessment questions

1. Antibody–antigen binding is based on _____ interactions.
   a. electrostatic (polar/nonpolar)
   b. hydrogen bonding
   c. ionic binding
   d. all of the above
2. How will increased temperature affect the reaction between antibodies and antigens?
   a. Increase rate, decrease binding affinity
   b. Decrease rate, increase binding affinity
   c. Increase rate, increase binding affinity
   d. Decrease rate, decrease binding affinity
3. An ELISA can be characterized as a:
   a. homogenous competitive binding immunoassay
   b. heterogenous competitive binding immunoassay
   c. heterogenous immunometric assay
   d. homogenous immunometric assay
4. Commonly used fluorescent labels include:
   a. Lucifer red
   b. FITC
   c. ethidium bromide
   d. all of the above
5. Immunoassays based on the EMIT detection technology usually utilize b-galactosidase as the enzyme label.
   a. True
   b. False
6. Heterophilic antibody interferences with immunoassays can occur when:
   a. patients have received antibody-based imaging reagents
   b. patients are exposed to mouse antigens by having mice in their living space
   c. patients have been given mouse monoclonal antibody-derived therapeutic agents
   d. all of the above
7. The hook effect in immunometric assays causes:
   a. falsely low patient results
   b. patient results to drop in a nonlinear fashion upon dilution
   c. reagent antibodies to bind irreversibly (i.e., "hook") to the antigen
   d. all of the above
8. Nonstandardized immunoassay kits are produced when:
   a. reagent antibodies from different manufacturers recognize different epitopes
   b. there is no primary reference material for the immunoassays
   c. neither A or B
   d. both A and B

Answers

1. d
2. a
3. c
4. b
5. b
6. d
7. a
8. d

Chapter 13

Nucleic acid analysis in the clinical laboratory

Sophie J. Deharvengt[1,2], Lauren M. Petersen[1,2], Hou-Sung Jung[1,2], and Gregory J. Tsongalis[1,2]

[1]*Department of Pathology and Laboratory Medicine. Dartmouth-Hitchcock Medical Center and Norris Cotton Cancer Center, Lebanon, NH, United States,* [2]*Geisel School of Medicine at Dartmouth, Hanover, NH, United States*

Learning objectives

After reading this chapter, the reader will be able to:

- Describe nucleic acid biochemistry and important features of deoxyribonucleic acid and ribonucleic acid.
- Understand the basis of Southern blotting and FISH.
- Explain the multiple uses of nucleic acid probes.
- List the steps involved in PCR.
- List the various chemistries used in qPCR.
- Describe the different platforms used in next-generation sequencing and their associated library preparation methods.

Introduction

The past decades have seen significant advances in the field of genomics. Major advances in sequence analysis and computing technology over the last three decades have led to the successful completion of the Human Genome Project [1], Human Microbiome Project [2], and countless other sequencing-based studies. This influx of information has significantly increased our understanding of human disease processes, leads to the development of new diagnostic testing algorithms, and is now pushing the boundaries of precision (or individualized) medicine. A multitude of laboratory tests now exists to study both deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). This chapter will focus on both the basic elements of how these tests work as well as how subsequent results are interpreted.

The biochemistry of deoxyribonucleic acid

DNA is a polymeric molecule composed of repeating nucleotide subunits. The order of nucleotide subunits contained in the linear sequence of this polymer represents all of the genetic information carried by a cell and defines the functional gene unit. The DNA structure is a double helix, composed of two polynucleotide strands that are coiled about one another in a spiral [3] (Fig. 13.1). There are four nucleotides that make up a DNA strand: adenine, guanine, cytosine, and thymine (Table 13.1). Adenine pairs with thymine and guanine pairs with cytosine. Each nucleotide is composed of (1) a phosphate group; (2) a pentose (five carbon atoms) sugar; and (3) a cyclic nitrogen-containing compound referred to as a base. The two-carbon nitrogen ring bases are purines (adenine and guanine), whereas the one-carbon nitrogen ring bases are referred to as pyrimidines (thymine and cytosine). Each polynucleotide strand is held together by phosphodiester bonds, linking adjacent deoxyribose moieties. The two polynucleotide strands are held together by a variety of noncovalent interactions including lipophilic interactions between the adjacent bases and the hydrogen bonding between the bases on opposite strands. The sugar-phosphate backbones of the two complementary strands are antiparallel and, thus, possess opposite chemical polarity. As one moves along the DNA double helix in one direction, the phosphodiester bonds in one strand will be oriented 5′-3′, while in the complementary strand, the phosphodiester bonds will be oriented 3′-5′. This configuration results in base pairs being stacked between the two chains perpendicular to the axis of the molecule.

Transcription is the first step of DNA-based gene expression, in which a specific portion of DNA is copied into RNA. RNA differs from DNA in that it is a single-stranded molecule and the nucleotide thymine is replaced with uracil. Another difference between RNA and DNA is that the nucleotides in RNA contain the sugar ribose instead of deoxyribose. There are multiple different types of RNA that play important roles in normal cellular

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00013-2

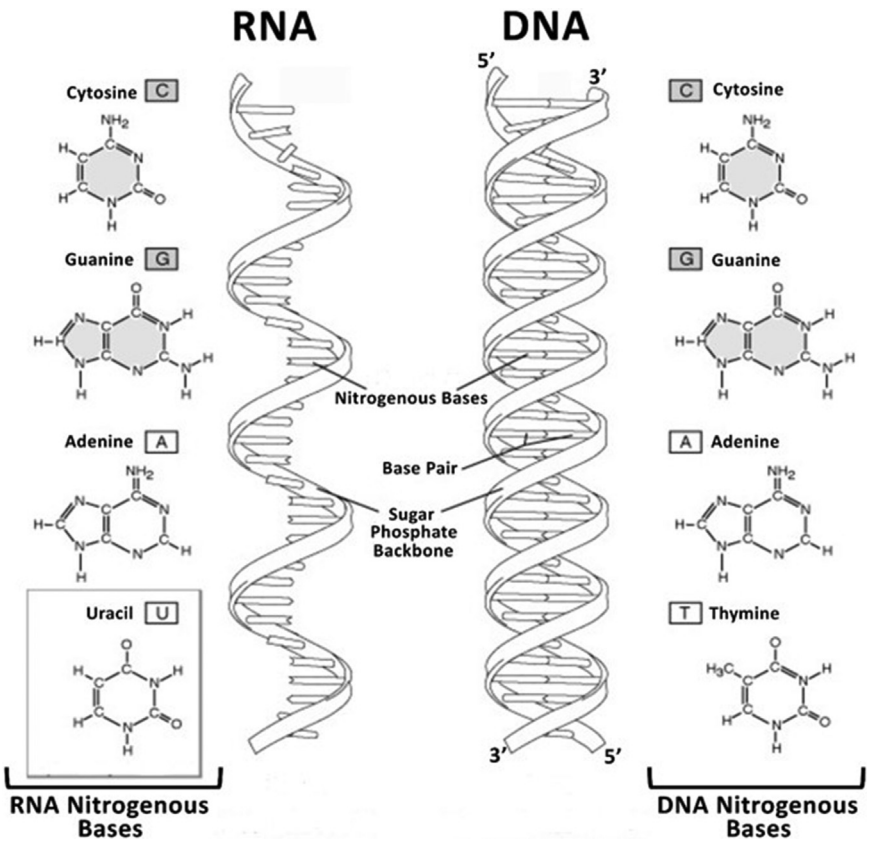


FIGURE 13.1 The structure of deoxyribonucleic acid.

TABLE 13.1 The nucleotides of DNA and RNA and their classification.

Nucleotide	Found in	Classification
Adenine	DNA and RNA	Purine
Guanine	DNA and RNA	Purine
Cytosine	DNA and RNA	Pyrimidine
Thymine	DNA	Pyrimidine
Uracil	RNA	Pyrimidine

DNA, Deoxyribonucleic acid; *RNA*, ribonucleic acid.

processes. The different types of RNA and the functions they perform are discussed in more detail in the next section.

The human genome

The genome of the typical human cell contains approximately 3×10^9 DNA base pairs that are subdivided into 46 discreet structural units, termed chromosomes. The human genome is diploid in nature, meaning the typical human cell contains 23 pairs of chromosomes. There are 22 pairs of autosomes and the sex chromosomes X and Y (females being XX and males being XY). The length and number of genes per chromosome vary greatly, but all together, the human genomes contain roughly 22,000 protein encoding genes. "Coding" DNA sequences give rise to all of the transcribed RNAs of the cell, including messenger RNA (mRNA), which encodes the amino acid sequence of a polypeptide. The majority of the human genome ($>98\%$) is made up of "noncoding" sequences, which have been suggested to function in DNA packaging, chromosome structure, chromatin organization within the nucleus, or in the regulation of gene expression. The majority of eukaryotic genes are comprised of coding regions, knowns as exons, which are broken up by specific noncoding sequences, called introns. During transcription, both exons and introns are copied into pre-mRNA; however, the process of RNA splicing removes introns. Exons are the portions of human genes that make up mature mRNA strands that are subsequently used for building proteins, a process known as translation. In addition to mRNA, many other types of RNAs exist in the cell and include ribosomal RNA (rRNA), transfer RNA (tRNA), and small nuclear RNA (snRNA). rRNAs form complexes with ribosomal proteins to make up ribosomes, which translate mRNA into amino acids. tRNAs are

responsible for transferring amino acids to ribosomes during translation. snRNAs are only found in eukaryotes and are involved in complexes that are important in RNA processing. In addition, many other types of RNAs exist that are integral for DNA replication, posttranscriptional modification, and protein synthesis.

Considerable variability in the human genome has been identified between individuals. There can be upward of 1.1−1.4 million single nucleotide polymorphisms (SNPs) distributed throughout the genome. Most SNPs have no direct consequences on the health of a person; however, some SNPs can result in major changes to the gene's function, causing developmental and health issues. Being able to detect and characterize SNPs has proven to be important in medicine. Genetic modifications can affect gene expression and can be stratified into different categories: (1) substitution (exchange of one nucleotide by another); (2) insertion (incorporation of extra nucleotides); (3) deletion (loss of a region of DNA); and (4) frameshifting (shift the way the sequence is read). While the majority of pathogenic SNPs occur within exons, intronic SNPS may also result in harmful effects, especially those that occur in the splice donor or acceptor sites, which can have negative functional effects on splicing.

Complementarity

In order to analyze the human or microbial genome, several techniques have been developed using the complementary base pairing property of DNA. Watson and Crick [4] described this homology, whereby the adenine on one strand of DNA binds to the thymine on the opposite strand, and likewise cytosine binds to guanine. This specificity results from the hydrogen bonding properties of the bases themselves. Adenine and thymine can form two hydrogen bonds, and guanine and cytosine can form three hydrogen bonds. The specificity of molecular interactions within the DNA molecule enables one to predict the sequence of nucleotides in one strand if the sequence of nucleotides in the complementary strand is known. This biochemical principle allows for the rapid design of synthetic primers and probes once the sequence of the target is known. A primer is a strand of short nucleic acid sequence (generally ranging from 10 to 30 bases) that is complementary to the target sequence and is used by the enzymatic DNA polymerase to initiate replication. A probe consists of a homologous single-stranded nucleic acid sequence that can hybridize to a target genomic sequence. Probes carry detection molecules that allow for the detection and visualization of the hybridized probe to its target. How probes and primers are designed and used in nucleic acid analysis will be described in the following sections of this chapter.

Southern blot

In molecular biology, hybridization probes are fragments of DNA or RNA, usually between 30 and 1000 bases, which hybridize to a specific target within the genome. Probes are designed and utilized in a number of techniques. Notably, one common characteristic of the utilization of probes is that, in all cases, the genomic sequence of the target must be known. Southern blotting is an early example demonstrating the use of molecular probes for DNA identification. Southern blotting, developed in the 1970s, was one of the first technologies utilized in clinical settings [5]. The overall methodology of a Southern blot is to target a specific sequence of DNA with a labeled probe so that it can be detected, usually by incorporating radioactivity or tagging the molecule with a fluorescent or chromogenic dye. The process begins by fragmenting genomic DNA into smaller pieces via the use of restriction enzymes. Restriction enzymes are bacterial nucleases that recognize specific sequential motifs and, subsequently, cleave DNA at these restriction sites. The small pieces of DNA are electrophoresed on an agarose gel for size discrimination [6]. Mutations, deletions, or rearrangements occurring within a gene may result in a disruption of the normal nucleotide sequence, possibly altering the number of restriction sites within the gene or altering the size of the restriction fragments produced. Such a change in the patterning of DNA fragments produced by enzymatic cleavage is referred to as a restriction fragment length polymorphism. Complete cleavage of the genomic sample is essential for Southern blotting, especially when single-copy genes are analyzed. Digestion of genomic DNA is confirmed by the presence of a continual ladder of fragmented DNA along with distinct bands, usually in the lower portion of the gel. The fragmented DNAs are subsequently denatured in an alkaline solution (typically with sodium hydroxide), transferred to a nylon solid support membrane via capillary action or vacuum transfer, and are bound permanently to the membrane by brief ultraviolet (UV) crosslinking or by incubation at 80°C. Once DNA has been immobilized to a membrane, specific sequences of interest can be detected by hybridization with a labeled, denatured, single-stranded probe that binds to its homologous or complementary sequence (Fig. 13.2A). Visualization of a radioactive or fluorescent labeled probe is accomplished by exposing the membrane to an X-ray film. If a chromogenic detection is used, then visualization will be accomplished by the development of color on the membrane itself.

The interpretation of a Southern blot can be relatively straightforward when one simply wants to identify the presence or absence of a gene sequence. Gene amplification, the presence of multiple copies of a gene sequence, can be viewed as a band that is more intense than the normal single-copy control gene sequence. Structural gene

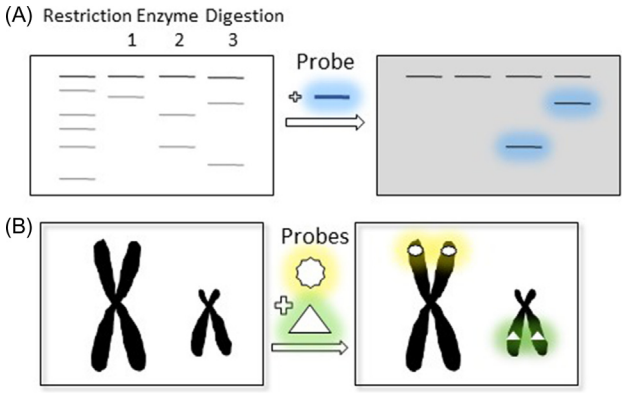


FIGURE 13.2 Schematic representation of nucleic acid probes in (A) Southern blotting and (B) fluorescence in situ hybridization.

alterations such as insertions, deletions, gene rearrangements, and point mutations in restriction enzyme recognition sites, can be detected by Southern blotting by observing banding patterns that are different from known control DNA samples.

Fluorescence in situ hybridization

Fluorescence in situ hybridization (FISH) is a molecular cytogenetic technique that has revolutionized the way chromosomes are examined [7]. Developed in the 1980s, FISH is used for examining the cellular reproduction cycle, specifically during interphase, to identify chromosomal abnormalities. It uses fluorescent probes that bind to specific targets on a chromosome with a high degree of sequence complementarity. Fluorescence microscopy is then used to identify where the fluorescent probe is bound to the chromosomes (Fig. 13.2B). The most significant development with FISH was the analysis of cells during interphase. Interphase analysis allows for cellular localization of target signals that can be associated with the morphologic phenotype. FISH is routinely used in the clinical laboratory to look for chromosomal abnormalities and gene mutations in individuals with certain diseases, such as Prader–Willi syndrome, Down syndrome, and cancer. However, the major limitation of FISH is that detection is limited to a specific abnormality for which probes are targeted.

FISH probes

A number of different kinds of probes have been developed for use in FISH. Whole chromosome probes (WCPs) are collections of small probes that bind to different sequences along the length of a chromosome. One can label all the WCPs specific to a single chromosome with one unique color, making it possible to identify when a piece of one chromosome is attached to the end of another chromosome (also known as a translocation). Locus-specific identifier (LSI) probes are used to determine if a certain gene is present in the genome and/or how many copies of that gene exist. Centromeric probes target the repetitive sequence regions of the chromosome centromeres. The common application of centromeric probes is to detect numerical chromosomal abnormalities and is referred to as chromosome enumeration probes (CEPs). Usually, a combination of an LSI probe and a centromeric probe that both target the same chromosome is used, with the CEP used as the internal control. In breast cancer, a polysomy 17-associated increase in *HER2* gene copy number is estimated using a combination of *HER2* LSI probe and a CEP 17 control probe. A multifold increase of the *HER2* gene on chromosome 17 over the chromosome copy number implies amplified expression of *HER2*, which has significant therapeutic implications. This is just one example of how FISH is used in the clinical laboratory.

The polymerase chain reaction

The polymerase chain reaction (PCR) was developed in the 1980s by Mullis and has had an unprecedented impact in the field of molecular diagnostics [8–10]. Mullis developed the PCR while working as a chemist at a biotechnology company called Cetus Corporation, first applying the technology to analyze mutations in sickle cell anemia. Mullis would go on to be awarded the Nobel Prize in Chemistry in 1993 for his work. Routinely used in clinical molecular diagnostic laboratories, PCR allows for the rapid amplification of target nucleic acid sequences to a level that can then be easily analyzed by other methods, such as chemiluminescence, fluorescence, and mass spectrometry. In doing so, PCR allows for both the qualitative and quantitative analyses of target sequences with high selectivity and sensitivity. PCR can be performed on virtually any sample type that contains enough biomass to extract DNA from. In the clinical laboratory setting, this includes blood, tissue, and formalin-fixed paraffin-embedded tissue, among others. This is advantageous if there is little starting material or in cases where the target is a pathogen derived from a patient as most of the extracted DNA will be of human origin and only a small portion obtained from an infectious organism. Numerous developments and improvements to the principles and applications of PCR have occurred since the 1980s, thus expanding how the technology may be used in both the research and clinical realms.

The process of polymerase chain reaction amplification

In a typical PCR, successive cycles are performed in which DNA polymerase copies a target DNA sequence

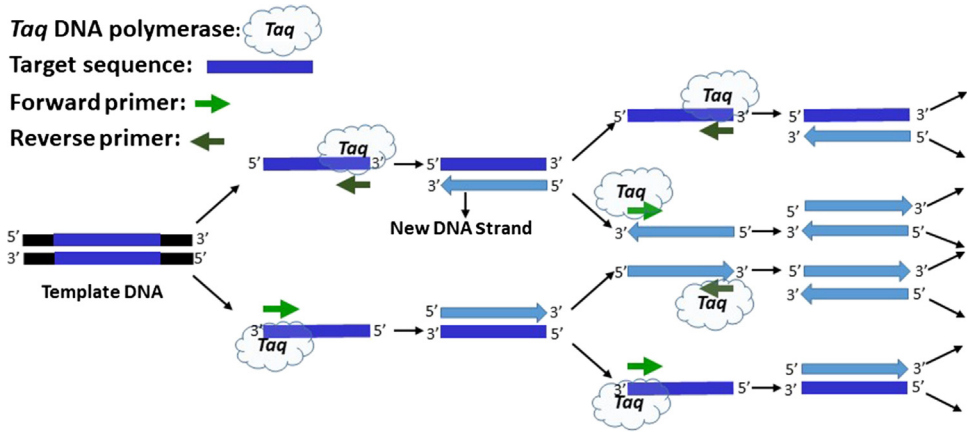


FIGURE 13.3 Schematic representation of the polymerase chain reaction.

from a template molecule in vitro (Fig. 13.3). The amplification products of each cycle provide new templates for the subsequent amplification. Thus the concentration of the target DNA sequence increases exponentially over the course of the PCR, increasing the analytical selectivity of the assay. The typical PCR mixture contains (1) a thermostable DNA polymerase (most commonly, Taq polymerase is used); (2) target-specific forward and reverse oligodeoxynucleotide primers; (3) deoxynucleotide triphosphates (dNTPs); (4) reaction buffer; and (5) a source of DNA template [genomic DNA or complementary DNA (cDNA)].

The active component of the PCR is a DNA polymerase enzyme that is required for DNA synthesis during the primer extension step. The contemporary PCR employs Taq DNA polymerase (isolated from *Thermus aquaticus*) [11]. Taq polymerase exhibits $5'\rightarrow3'$ polymerase activity, $5'\rightarrow3'$ exonuclease activity, thermostability, and optimum performance at 70°C–80°C. Temperature, pH, and ion concentrations (Mg^{2+}) can influence the activity of Taq polymerase. The half-life of Taq activity at 95°C is approximately 40–60 minutes, and extremely high denaturation temperatures (>97°C) will significantly reduce its active lifetime. Since time at temperature represents the critical parameter for maintenance of Taq activity, lowering of the denaturation temperature or reduction in the denaturation time can prolong the activity of the enzyme during a PCR. The typical PCR is carried out in a buffer (usually Tris-Cl) of pH 8.3. Taq polymerase requires divalent cations in the form of Mg^{2+}. Mg^{2+} binds to DNA, primers, and nucleotides contained in the amplification reaction. Lower divalent cation (Mg^{2+}) concentrations decrease the rate of dissociation of the enzyme from the template by stabilizing the enzyme–nucleic acid interaction. Most PCR mixtures contain at least 1.5 mM $MgCl_2$. However, a $MgCl_2$ titration is recommended for any new template–primer combination to determine the optional concentration. While Taq DNA polymerase is ideal for routine PCR, there are several other thermostable DNA polymerases with unique qualities, which make them useful for special PCR applications such as amplification of long pieces of DNA or high-fidelity amplification [12].

The target sequence is defined by the specificity of the oligodeoxynucleotide primers that anneal to complementary sequences on opposite template strands flanking the region of interest (ROI). During the PCR, these primers are extended in the $5'\rightarrow3'$ direction by the polymerase enzyme to yield overlapping copies of the original template. Each cycle of the PCR proceeds through three distinct phases: (1) denaturation; (2) primer annealing; and (3) primer elongation or extension. The denaturation step consists of destabilizing the double-stranded DNA (dsDNA) template by disrupting the hydrogen bonds between the complementary bases; this process allows access for the primer to hybridize. The denaturation is typically accomplished by incubating samples for up to 1 minute at >94°C. Primer annealing is accomplished at a temperature that is specific for the PCR primers and conditions employed; these conditions may be empirically determined during assay development. The optimal annealing temperature for a given oligonucleotide primer set is very important for PCR specificity. Generally, the annealing temperature used should be no more than 5°C below or higher than the lower primer melting temperature (Tm) and typically ranges from 55°C to 65°C. The Tm of an oligonucleotide primer can be calculated using a simplified formula that is generally valid for primers that are 18–24 bp in length. This formula is as follows: $Tm = 69.3 + 0.41(\%G + C) - (650/L)$, where L is the primer length in bases [13]. During the annealing step, primers hybridize to the complementary target sequence contained within the denatured single-stranded template. This step produces a starting point for the polymerase enzyme to catalyze the incorporation of dNTPs in a DNA-directed DNA synthesis reaction. However, due to the high concentration of the primers in the reaction, weak interactions can occur inducing nontarget amplification.

This phenomenon is known to have primer−dimer formation.

The primer extension/elongation step is accomplished at approximately 72°C. The actual times used for each cycle will vary from 15 seconds to 1−2 minutes depending on both the DNA polymerase used and on the length of the DNA region to be amplified (also known as the amplicon).

Amplification of target sequences is accomplished through repetition of these aforementioned steps (denaturation, annealing, and elongation) for 25−50 cycles. The exact number of cycles necessary to produce an amplicon that can be detected will depend on the starting concentration of the DNA template, size of amplicon, as well as downstream detection method. By the end of the third cycle of amplification, a new double-stranded molecule is formed in which both the 5′-end and 3′-end coincide exactly with the primers employed. Since the copy number theoretically doubles after each successive cycle, an exponential increase of 2^n (n being the total number of cycles) is accomplished during the complete reaction. Accumulation of amplicons corresponding to the target sequence eventually levels off and reaches a plateau, due to the exhaustion of PCR reagents and loss of activity of the polymerase. The initial number of target sequences contained within the template sample, the efficiency of primer extension, and the number of PCR cycles performed determine the upper limits of amplification.

There are many advantages and disadvantages to PCR. In addition to being sensitive and specific, starting material does not have to be genomic DNA but can be cDNA, which is discussed in the next section. The size of amplicons can also vary according to the result needed for clinical use. The size of the amplicons typically ranges from 50 to 1000 nucleotides; however, PCR with the appropriate polymerase can produce amplicons above 1000 bp. A drawback of PCR is the error rate. While Taq polymerase has a reported error rate of $1-20 \times 10^{-5}$, other high-fidelity polymerases are reported to have higher rates of accuracy [14]. Even so, if an incorrect base is incorporated into a newly synthesized strand in the early steps of PCR, that error is carried through the rest of the amplification process. This can lead to erroneous results, which can have negative consequences, especially in genotyping PCRs or when the PCR product is then used for sequencing. The chance of incurring errors increases in PCRs with long amplicons.

Reverse transcriptase polymerase chain reaction

In order to expand the power of PCR amplification to the transcriptome level, RNA is reverse-transcribed into cDNA in a reverse transcription reaction [15]. Reverse transcriptase (RT; retroviral RNA-directed DNA polymerase) is the enzyme used to catalyze cDNA synthesis. The RT reaction consists of five components: (1) a single cDNA synthesis primer; (2) an appropriate RT buffer; (3) dNTPs; (4) RNA template (total RNA or mRNA); and (5) RT enzyme. There are several commercially available RT enzyme preparations that can be used in standard reverse transcriptase polymerase chain reaction (RT-PCR) applications. These include RT from Moloney murine leukemia virus and avian myeloblastosis virus. More recently, recombinant derivatives of these RT enzymes have become available, offering advantages over the native enzymes. Advanced enzyme preparations like these produce the highest yields and confer high specificity when gene-specific primers are employed.

RT-PCR is a commonly employed method for analysis of RNA transcripts, especially for measuring low abundance species or working with limited amounts of starting material. Classic blotting and solution hybridization assays require much more RNA for analysis and lack the speed and ease of technique afforded by PCR-based applications. RT-PCR couples the tremendous DNA amplification capabilities of the PCR with the ability to reverse transcribe small quantities of RNA into cDNA. Other advantages of RT-PCR include versatility, sensitivity, and the ability to compare multiple samples simultaneously.

Analysis of polymerase chain reaction amplicons

There are numerous methods to analyze PCR products, and the method of choice depends on the type of information desired. In most cases, amplicons can be analyzed by standard agarose gel electrophoresis, which effectively separates DNA products over a wide range of sizes [from 100 bp to >25 kilobases (kb)]. Size markers can be electrophoresed on the gel to allow size determination. PCR products from 200 to 2000 bp can be separated quickly on a 1.2%−2.0% agarose gel. When greater resolution or separation power is required, such as in the analysis of very small amplicons (<100 bp), polyacrylamide gel electrophoresis (PAGE) is the method of choice. In both cases, a fluorescent dye that intercalates into the DNA can be added to the gel during preparation; alternatively, the dye can be added following electrophoretic separation for gel staining. Ethidium bromide (EtBr) is one such fluorescent dye that has been used for visualizing DNA on gels. Because EtBr is an intercalating agent, it can be potentially genotoxic. However, alternatives that are both safer and better performing have been developed and include Green-Glo and SYBR Safe. Regardless of the type of fluorescent dye used, DNA fragments are easily visualized as bands by UV-illumination.

If the clinical assay requires measuring the amplicon size with very high resolution (e.g., sequencing or fragment analysis), laboratories have turned to capillary electrophoresis (CE) systems [16]. Unlike standard SDS-PAGE separation, CE systems can provide molecule detection after separation through a light source and a detector. In addition, the system can utilize high voltages without overheating the samples. The CE was essential for the completion of the Human Genome Project. Although CE's role in de novo genome sequencing has decreased with the replacement of Sanger sequencing by the next-generation sequencing (NGS; discussed later in this chapter), CE is still used to improve quality control in the NGS by determining the amplicon size. CE is also widely used for DNA sizing in other methodologies, for example, in assigning short tandem repeat (STRs) and variable number tandem repeats for human identification or detecting pathogen biomarkers [17]. CE can also be used in multiplex ligation-dependent probe amplification assays, commonly used for the detection of unusual copy number changes (insertions or deletions) of human genes for the diagnosis of several genetic diseases whose pathogenesis is related to the copy number variation (CNV) of specific genes. A limit of CE is the requirement of expensive equipment but, once acquired, it has the advantages of low operating costs, small sample input, and high resolving power.

In some cases, the desired information resulting from a PCR analysis can be obtained through a simple analytical gel separation, as just discussed. However, in other instances, additional information is required. PCR products can be directly sequenced after purification. In addition, they can be cloned and used for the construction of molecular probes, in mutation analysis, in in vitro mutagenesis, or in gene expression studies.

Multiplex polymerase chain reaction

Another important milestone in the evolution of PCR involves the expansion of assay multiplexing capabilities [18]. Multiplex PCR is a powerful technique that enables the amplification of two or more products in parallel in a single reaction tube. It employs different primer pairs in the same reaction for simultaneous amplification of multiple targets. Multiplex PCR is used widely in genotyping applications and multiple areas of DNA testing in research, forensic, and diagnostic laboratories. Applications include gene expression and deletion analysis, SNP genotyping, forensic identity testing (e.g., STR typing), and pathogen detection. Quantification of multiple genes in a single reaction also reduces reagent costs, conserves precious sample material, and allows increased throughput.

Real-time polymerase chain reaction

A significant advance in the PCR methodology was the development of real-time PCR (qPCR) assays in the 1990s. This technology allows for simultaneous real-time detection of PCR products, as they are amplified with the use of fluorescent detection (Fig. 13.4) [19]. Various detection chemistries for qPCR were rapidly introduced and replaced many of the older detection methods (including gel electrophoresis, allele-specific oligonucleotide blots, and others). The fluorescent detection systems used in qPCR offer numerous advantages over most molecular technologies including closed-tube amplifications, decreased concern for contamination, the ability to perform qualitative or quantitative analysis, and the ability to perform variant screening with melt curve analysis.

The reason qPCR is so rapid is that it combines the amplification steps of traditional PCR with simultaneous detection steps that do not require post-PCR manipulation, because the PCR is monitored directly within the reaction vessel. In qPCR, the exponential phase of PCR is monitored, as it occurs using fluorescently labeled molecules [20,21]. During the exponential phase, the amount of PCR product present in the reaction vessel is directly proportional to the amount of emitted fluorescence and the amount of initial target sequence. The higher the copies of input DNA, the fewer the amplification cycles are needed to achieve a significant increase in fluorescence, allowing these reactions to be quantitative, resulting in the nomenclature of qPCR.

There are several types of detection chemistries for qPCR: those which use intercalating DNA binding dyes such as SYBR Green I and those that use various oligonucleotide probes with attached fluorophores [22]. Intercalating DNA binding dyes allow for the simple

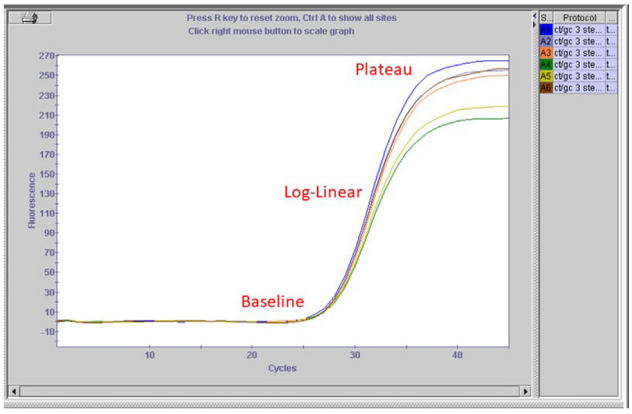


FIGURE 13.4 A real-time polymerase chain reaction amplification curve showing the baseline, log-linear, and plateau phases of the reaction. This set of curves was generated using the Smartcycler II (Cepheid, Sunnyvale, CA).

determination of the presence or absence of an amplicon. SYBR Green I, like EtBr, is a dye that emits fluorescence when it is bound to dsDNA. During the PCR, there is an increase in the copy number of the amplicon, as well as a simultaneous increase in the amount of intercalated SYBR Green I. This will then increase the level of emitted fluorescence in direct proportion to the copy number. One disadvantage to these types of dyes is that they are nonspecific and will bind to any dsDNA. To increase further detection specificity, primer-probes are used, and they are discussed in the next section.

Probes for real-time polymerase chain reaction

Probes for qPCR can be divided into two categories that are based on the fates of probes during PCR: a hydrolysis-type probe and a hybridization-type probe. The hydrolysis-type probes have a reporter fluorophore at one end of the probe and a corresponding quencher at the opposite end (Fig. 13.5). The reporter and the quencher are so close that excited energy from the reporter is transferred to the quencher by Förster resonance energy transfer (FRET); consequently, no fluorescence is emitted as the excitation energy is absorbed by the quencher molecule. When the probe hybridizes to its newly synthesized target sequences, the reporter fluorophore is released from the probe through the 5′→3′-exconuclease activity of the DNA polymerase during qPCR. The released fluorophores emit signals when they are excited by a specific wavelength of light, and the amount of the signal is proportional to the amount of PCR product generated. The most common example is the TaqMan probe (Fig. 13.5A).

There are several methodologies behind how hybridization-type probes work during qPCR [23]. The dual hybridization probe strategy utilizes two probes simultaneously, which are designed to hybridize adjacent to each other on the same target sequence. The first probe is labeled on the 3′ end with a donor fluorophore and the second probe is labeled on the 5′ end with an acceptor fluorophore. When these probes hybridize to the target sequence and are in close proximity, excitation of the donor fluorophore results in a transfer of energy to the acceptor fluorophore. This FRET process results in the emission of light at a wavelength specific to the acceptor fluorophore (Fig. 13.5B). By measuring the signals emitted at the wavelength of the donor and acceptor fluorophores, the amount of FRET can be determined by the qPCR instrument. An example is the HybProbes or FRET probes. One of the drawbacks to a hybridization probe-based approach is that two probes need to be designed instead of a single probe.

Another hybridization-type probe strategy takes advantage of the hairpin structure in which 5′ and 3′ ends of probes are close enough to allow fluorescence quenching in the absence of target sequences. Following the probe's hybridization to target sequences, the hairpin structure melts. This provides the reporter with enough distance from the quencher for fluorescence emission (Fig. 13.5C). This type of probe is known as the hairpin probe or the molecular beacon probe (HyBeacon). This probe is highly specific, because the hairpin and hybridized structures are interchangeable so that even an SNP in the target sequence results in the hairpin structure and not hybridization.

Alternatively, HyBeacon probes provide a method for fluorescence-based sequence detection, allele discrimination,

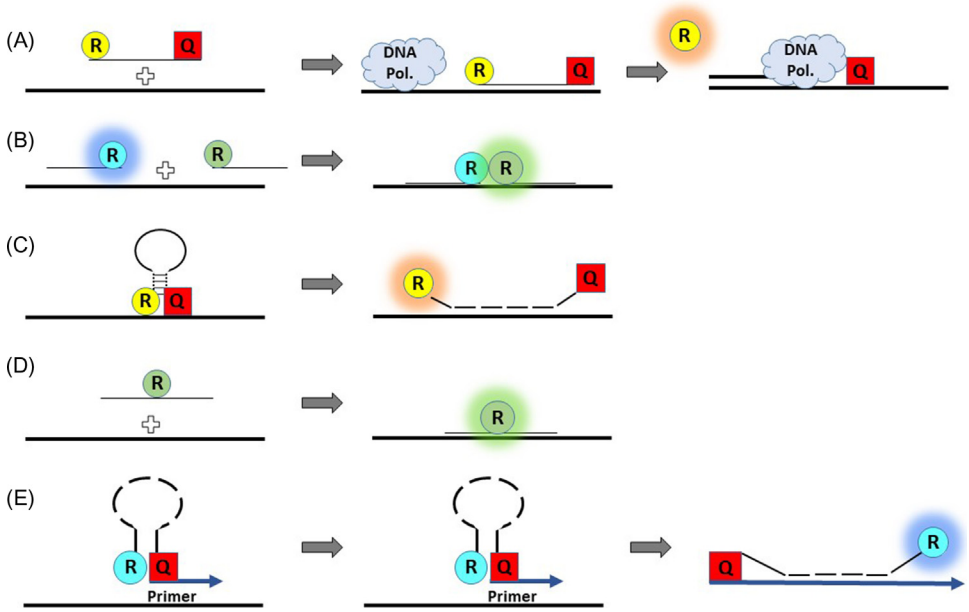


FIGURE 13.5 Schematic diagrams of probes used in real-time polymerase chain reaction. (A) TaqMan probes. (B) HybProbe or Förster resonance energy transfer probes. (C) Molecular beacon probe. (D) HyBeacons probe. (E) Primer-probe. The glowing images represent emitted fluorescence signals. *DNA Pol*, DNA polymerase; *Q*, quencher of fluorescence from fluorophore (R); *R*, reporter fluorophore.

and DNA quantification [24]. In the case of HyBeacon probes, a reporter fluorophore is used without any quencher (Fig. 13.5D). When the reporter is attached to the middle of a probe sequence instead of at the end of the probe (like the Taqman or FRET probes), fluorescence emission rates are considerably lower in a single-stranded state than a double-stranded state. This is because the probe alone emits weak fluorescence signals and the probe emits stronger fluorescence signals when it hybridizes to a target sequence. Design of the HyBeacon probe is simple, and it is inexpensive, because only one fluorophore is required.

Primer-probes are a group of probes used in qPCR and harbor both probe and primer capabilities in a single oligonucleotide molecule [21]. Primer-probes typically have an attached fluorophore molecule and a hairpin structure, which places the reporter fluorophore close to an attached quencher to reduce the background emission signal from the reporter (Fig. 13.5E). Primer-probes are the aforementioned molecular beacon probe combined with primer oligonucleotides. When the probe part of a primer-probe hybridizes to its adjacent, newly synthesized target sequence, the reporter emits fluorescence signals and the amount of signal is proportional to the amount of target sequence. The advantage of primer-probes is the identification of amplified product if there are shared probe-hybridization sequences in multiplex PCR assays. Other advantages are inhibition of primer−dimer formation and prevention of nonspecific PCR signal detection. Disadvantages include cost of the primer-probe and designing probes close to the primer sites. Examples include Scorpions and Amplifluour primer-probes.

Post real-time polymerase chain reaction analysis

There are two main post-qPCR analyses: one using the amplification curve and the other using a melting, or denaturation, curve. Amplification curve analysis allows for absolute and relative quantitation of the amplicon concentration, during each amplification cycle. The approach is to set a quantification threshold and to determine the quantification cycle Cq [the number of cycles required to attend this threshold, also known as cycle threshold (Ct)]. The threshold should be set in the exponential phase of the PCR reaction. The fluorescent baseline for the amplification plot is a crucial step in qPCR data analysis, and this is typically set at an early point in amplification. Quantification is either obtained by using several different samples of known concentration or a standard, in order to construct a standard curve or to quantify using a comparative Cq method. The comparative Cq method uses the formula $2^{-\Delta\Delta Ct}$ in order to calculate the relative gene expression (relative quantitation) fold change relative to reference samples (such as an untreated sample or endogenous control or housekeeping genes). The comparative Ct method demands that the PCR efficiency of the biological sample and amplification of the reference are approximately equal. Absolute quantitation of unknown samples can only be obtained by using a standard curve made of standard with known absolute quantities.

Melting curve analysis is based on the fundamental property of dsDNA denaturation with heat [25]. This separation can be monitored with the use of fluorescent dyes that specifically fluoresce when bound to dsDNA. As the temperature is raised, the double strand begins to dissociate, releasing the dye and leading to a decrease in the fluorescence intensity. Data generated are represented by a curve of fluorescence intensity versus temperature. The point at which 50% of the DNA is in the double-stranded state is called the melting temperature (Tm) and represents the peak of the derivative of the melting curve (Fig. 13.6). In the presence of saturating concentrations of the DNA binding dyes, a specific amplicon sequence will present with a specific melting profile containing a specific Tm and a melting curve shape. This distinctive melting curve can be used to detect DNA sequence variations in the amplicon without the need for any post-PCR processing. The development of both high-resolution qPCR instruments and new saturating DNA dyes allows for a more accurate assessment of sequence variations based on melting analysis. High-resolution melting analysis (HRMA) can discriminate DNA sequences based on their

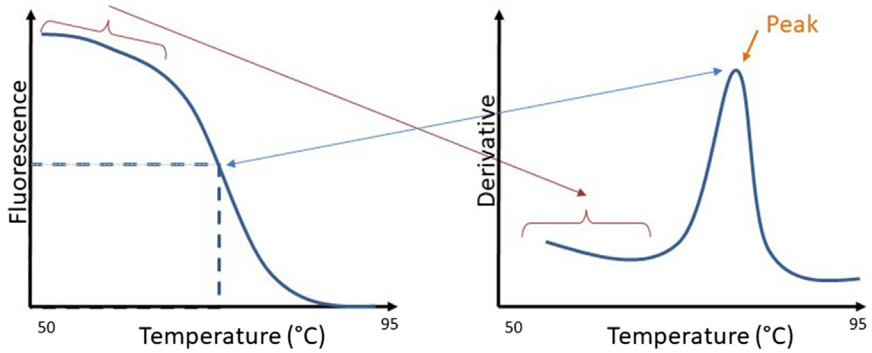


FIGURE 13.6 Melting curve output from real-time polymerase chain reaction.

composition, length, guanine-cytosine (GC) content, or strand complementarity. It can be used for mutation scanning experiments, methylation studies, and genotyping. The advantages of this method are its ease of use, high sensitivity, high specificity, and low cost. These characteristics make HRMA ideal for use in routine diagnostic settings [26]. However, the accuracy of the technique depends on appropriate instrumentation, saturation dyes, and software for analysis, which often introduces variability among clinical diagnostic laboratories.

Multiplexing with real-time polymerase chain reaction

The use of probes in qPCR allows for determination of more than one genotype in a single qPCR reaction. Modern qPCR instruments harbor up to five sets of excitation and emission filters for labeled fluorescence dyes [27]. When there are more than five known genotypes, the solution is to utilize independent, or separate, PCR assays. The introduction of multiplex technologies allowed for the detection of multiple variants in the same or separate genes simultaneously and in a single reaction tube.

Multiplexed genotyping often requires two probes per target molecule. One probe identifies the wild type or normal sequence, and the other detects the specific variant of interest. Typically, the various types of probes used are labeled with different fluorophores. Depending on the fluorescent signal detected, the genotype can be determined as either homozygous normal, heterozygous, or homozygous mutant.

The Idylla assays use a qPCR chemistry based on the PlexPrime and PlexZyme (also known as MNAzyme) technologies that allow multiplexed variant detection with high sensitivity and specificity [28]. With this technology, each primer is designed to have a 5′ target-recognition region, a short 3′ target-specific sequence complementary to the mutation of interest, and a distinct insert sequence that is mismatched (MM) to the target. This results in the production of allele-specific amplicons that are detected in real time by allele-specific PlexZyme enzymes and a universal fluorescent probe, allowing for the detection of a broad range of mutations in a single reaction. This system can be used for the rapid detection of somatic variants [29].

Multicolor melt curve analysis for high risk human papillomavirus (hrHPV) detection is a PCR-based DNA amplification method that enables simultaneous identification of amplified target DNA using melting temperatures of double-stranded targets from different hrHPV types.

Multicolor melting curve analysis is another way to use the potential of qPCR by combining the use of dual-labeled probes with high-resolution melting curve analysis. Such a system can be used for the rapid detection of human papilloma virus (HPV) [30]. In the assay developed by QuanDx, 14 different high-risk HPV genotypes are tested for within a single reaction because of high-resolution melting curve analysis (www.quandx.com).

Digital polymerase chain reaction

Digital PCR (dPCR) is a modification of conventional PCR methods that can be used to quantify directly and amplify clonally nucleic acids strands [31]. qPCR enables relative quantification of target sequences but not absolute quantification due to the issue of amplification bias and the requirement of standard references that can differ between the laboratories. dPCR offers significant benefits over other PCR methods by providing an absolute count of target DNA copies per input sample without the need for running standard curves. That makes this technique ideal for nucleic acid quantification, rare mutation detection, pathogen detection (viral load analysis and microbial quantification), CNV detection, microRNA expression analysis, NGS, single-cell gene expression analysis, and chromosome abnormality detection.

dPCR is derived from the digital readout, which utilizes numerical digits to identify “presence” or “absence” in single reactions. dPCR is performed by diluting target nucleic acid across a large number of reactions (termed partitions or fractions), which statistically results in one or zero molecules per reaction prior to amplification. A strong sample partitioning enables the reliable measurement of small fold differences in target DNA sequence copy numbers among samples. In addition, the creation of individual reaction compartments avoids the effects of cross-contamination and achieves absolute target quantification in each sample. However, there is always the possibility that more than one molecule will be present per reaction, thus requiring the need to use statistical methods, such as the Poisson distribution or negative binomial distributions.

There are four main ways to partition samples in dPCR: droplet-based sample dispersion (oil−water emulsions), microwell-based sample dispersion, channel-based sample dispersion, and printing-based sample dispersion [32]. The primary dPCR method is the droplet dPCR (ddPCR). After partitioning, the next step consists of thermal cycling amplification or isothermal amplification (Fig. 13.7). Then, each compartment is evaluated individually. This removes PCR efficiency bias from amplification and decreases error rates, thus enabling the detection of small fold differences. In addition, evaluation of an individual compartment in ddPCR reduces the background signal by increasing the signal-to-noise ratio, thus significantly improving the detection sensitivity. Fluorescent probes, such as the TaqMan hydrolysis probes or an intercalating dye, are generally used to detect the

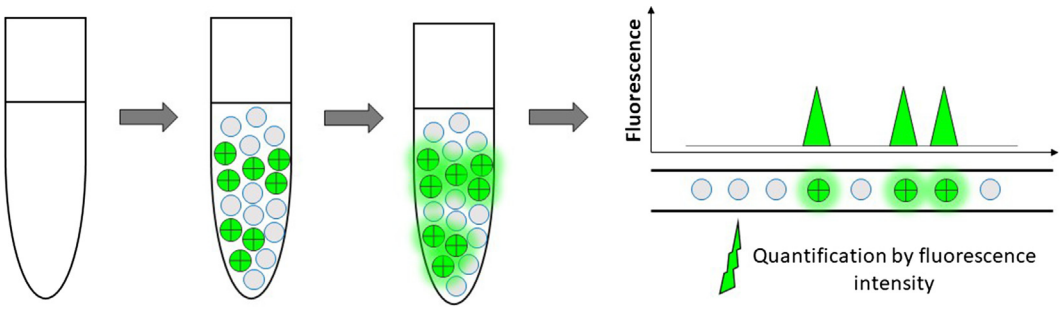


FIGURE 13.7 Overview of droplet digital polymerase chain reaction.

amplification products. A single template reaction yields a positive signal, which can be differentiated and eliminated from the fluorescence intensity generated by multiple target reactions. The ddPCR quantification process is droplet-based fluorescence signal counting, a method that determines the number of droplets that have different fluorescence intensities, similar to the principle of flow cytometry.

dPCR encounters the same challenges as PCR and qPCR in terms of experimental design and the need of internal controls. Though dPCR offers the potential to perform precise low abundance quantification, it is imperative to measure sources of false-positive signals due to low-level contamination and/or nontarget amplification (caused by mispriming and/or primer dimers) and may lead to overestimation. Conversely, PCR inhibition can lead to underestimation due to the fact that dPCR is only able to count positive partitions. This problem is also encountered when detecting RNA by RT dPCR. The use of an internal positive control, such as the housekeeping gene beta-actin, is highly recommended when measuring clinical samples, especially when reporting negative results. The optimal internal positive control gene will depend on the PCR conditions used for the genes targeted in the assay.

Digital polymerase chain reaction and multiplexing

Similar to the multiplexing applications with standard qPCR, it is also possible to duplex or multiplex with dPCR [33]. Competing duplex dPCR reactions can be utilized in which the same primer set is used but includes two probes, with each probe designed slightly differently in order to detect changes in a single targeted sequence. Each probe would have a unique reporter dye. Duplex or multiplex (if using more than two unique probes) reactions using this approach would be for used for detection SNPs, single nucleotide variants, and small insertions or deletions. In contrast, if using a noncompeting duplex or multiplex dPCR approach, each probe would have its own corresponding forward and reverse primer pair and each set of primers and probe would target a different nucleic acid sequence.

One of the current limitations in multiplex dPCR can be the number of available optic channels. Although newer instruments such as the Fluidigm BioMark and EP1 systems offer four optical channels for detections, many instruments only offer two channels for detection. Nevertheless, there are ways to apply higher order multiplexing to these two optic channel instruments by measuring the end-point fluorescence amplitude. This is possible with dPCR due to the advances in partitioning fluidics that enable a reaction to be subdivided into an increasing number of partitions. Each partition is a function of probe-dye conjugation and mixing, probe and primer concentrations, and the type and concentration of targets that are present. Thresholds must be determined for each partition, and appropriate software is required to analyze the data, which is often represented as a 2-D scatter plot. Clinical diagnostic assays such as genotyping and gene copy number can be determined using this approach.

Microarrays

A DNA microarray is another important clinical diagnostic method to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome (Fig. 13.8A) [34]. In contrast to the assays already discussed, microarrays can analyze large portions of an organism's genome due to the use of thousands or millions of different probes simultaneously. There are various approaches to perform microarrays but the basis of the microarray technology generally corresponds to a collection of microscopic features (DNA "spots") arranged on a chip in a grid-like format. These features can be probed with target molecules to produce either quantitative (gene expression) or qualitative (diagnostic) data. The spots correspond to the probe (e.g., DNA sequence) immobilized to a solid surface (i.e., membrane or beads). The nature of an array is such that it requires the location of every probe to be documented. This then allows for the identification of probes that have located complementary sequences/targets in the sample.

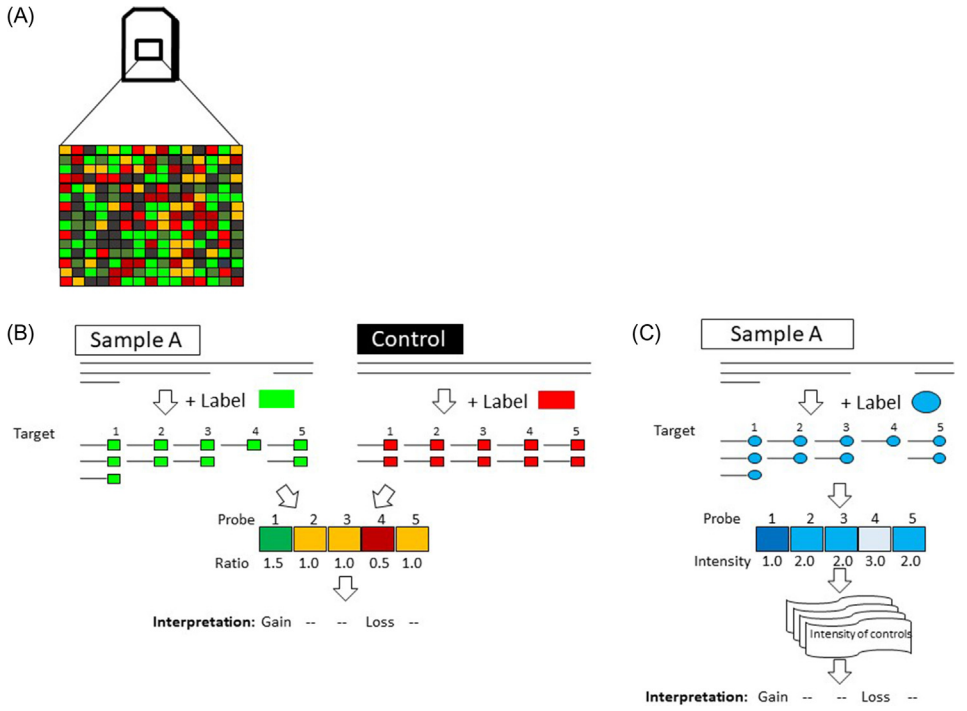


FIGURE 13.8 Microarray experiment design for copy number variants. (A) Overview of microarray. (B) Two-color microarray experiments. (C) One-color microarray experiments. In the two-color experiment (B), test sample and control are labeled with two different fluorescence dyes, displayed with white and black squares. Labeled targets are hybridized in a single microarray, and ratios of two fluorescence signals at each probe position are calculated to determine genome aberrations. In the one-color experiment (A), only testing sample is labeled and hybridized in a microarray. Signal intensities are compared with intensity data of a collection of controls to diagnose copy number variations.

The hybridization between the probe and the labeled target is detected and quantified using fluorophore, silver, or chemiluminescence methods. Due to technological advances and the variable experimental details (including probe length and synthesis, density of features, and the solid surface), the spectrum of microarrays is significant. Microarrays can be distinguished based on characteristics such as the nature of the probe (e.g., DNA) and the type of microarray.

Four major categories of microarray are available. The first category includes the printed arrays that use the process of "printing" or spotting of the probes onto the microarray surface, (glass microscope slide). The second category regroups the in situ-synthesized arrays that are high-density microarrays, such as GeneChips (ThermoFisher Scientific, Waltham, MA) and Sure Print (Agilent Technologies, Santa Clara, CA). The third category embodies high-density bead arrays (Infinium HD assay; Illumina, San Diego, CA), that uses silica beads as a support for the probes (i.e., specific oligonucleotides). The beads are randomly assembled and held in microwells on patterned substrates, allowing a uniform spacing. Each sample is fluorescently labeled and can be identified using distinctive indexes (IllumiCodes), which allows for multiplexing. The indexes are also used to map the location of the sample on the substrate. The fourth category pertains to suspension bead arrays, which are based on the use of spectrally distinctive beads as the solid support and the xTAG technology as the probe (Luminex Corporation, Madison, WI). The sample to be analyzed is mixed with unique oligonucleotide sequences (xTAG) specific to each target. To prevent cross-hybridization with any target sequence, the xTAG technology uses only three, instead of four, bases [thymine (T), adenine (A), and guanine (G)] in the antitag and tag sequences, hence the name of the xTAG technology. Binding with the TAG allows for allele-specific primer extension, during which biotin-labeled dCTPs are incorporated. The products are then mixed with color-coded beads. A specific antitag sequence is attached to a bead of specific color, and the tag sequences are hybridized to antitag sequences. The subsequent detection of a fluorescent reporter indicates that probe-target DNA hybridization has been accomplished using a special benchtop flow cytometer. An example of how this technology is used clinically is the laboratory developed test (LDT) Luminex cystic fibrosis genotyping assay, which can detect up to 71 mutations in the cystic fibrosis transmembrane ragulator gene in a single tube reaction (www.luminexcorp.com).

Microarray technologies also allow for CNV across the entire genome, which is far beyond the analysis capacity of qPCR or multiplex genotyping [35]. Microarrays for CNV identification, called array comparative genome hybridization (aCGH), utilize diverse types of probes [36]. In early aCGH approaches, BAC clones of approximately 100 kb were used as microarray probe and made it possible to cover the whole genome. However, the resolution was too low to detect small variations and to define break points. To improve resolution, cDNAs of 1−2 kb were developed for microarray probes. Limitations of cDNA probes, however, include lower

genome coverage and poor hybridization efficiencies when targets are genomic DNAs. To help resolve these technical issues, PCR products of about 1 kb were used as probes after being directly amplified from genomic DNA. Both PCR and cDNA probes, however, still have limitations with low resolution and technical difficulties of spotting these clones on limited space on a slide. To achieve higher resolution, oligonucleotides of up to 70 base pairs are used as microarray probes. Either ink-jet technology or photolithography generates over four times as many probes on arrays as older spotting technologies.

CNV microarrays are further divided based on experimental designs: two-color and one-color experiments. In the one-color experiment, biotin-labeled genomic DNA from a single patient hybridizes to probes (Fig. 13.8C), and then, streptavidin–phycoerythrin complexes convert the biotin label into a fluorescence signal. The signal intensities are measured as absolute values. Signal intensities at each probe are compared with a collection of controls to determine CNV. Absolute values from the one-color experiments can be directly compared with other experiments after intensity value normalization. In the two-color experiment, genomic DNA from a patient and a control are labeled with different fluorophores, such as Cy3 and Cy5 (Fig. 13.8B). These labeled targets hybridize to probes in a single array, and signals from Cy3 and Cy5 at each probe are converted into Cy5/Cy3 ratios. Statistically significant deviation in the signal ratio from expected distribution is interpreted as CNV. Ratio values from the two-color experiments can be compared only among experiments using the same control. Compared with one-color experiments, signal variation per probe is reduced in two-color experiments because of identical hybridization conditions. However, differences in labeling efficiency between Cy3 and Cy5 and in hybridization efficiencies between Cy3- and Cy5-labeled targets can influence fluorescence intensities and lead to inaccurate gene expression results [37]. To control for this, replicate microarrays can be performed in which the opposite dye orientations are used to control for gene-specific dye bias.

Microarrays are also widely used for identifying genome-wide single nucleotide polymorphisms (SNPs). There are two types of arrays for SNP genotyping: resequencing microarrays and genotyping microarrays. The biggest difference between the two approaches is in probe design (Fig. 13.9). In the resequencing microarrays, all four possible SNP variants (A, T, G, and C) are positioned at the center of the oligonucleotide probes with the remainder of the sequence being identical (Fig. 13.9A). Following hybridization, an SNP is detected by comparing the hybridization intensity among the four probes. Probes with the highest intensity identify the nucleotide in the middle of probe sequence. Resequencing microarrays can determine the identity of SNPs at any sequence variation position. However, there is no system for monitoring cross-hybridization, and therefore results may contain false positives in the cases where probes are binding to similar, but not identical sequences [38].

Genotyping microarrays have probes that hybridize to known allelic sequences (Fig. 13.9B). To increase the specificity of SNP calling, probes are designed to hybridize to target sequences at different levels, that is, a probe harbors an SNP in the middle of the oligonucleotide sequence, and other probes are shifted either upstream or downstream of the SNP position [39]. Accordingly, each probe is expected to hybridize to its target with a different intensity to determine the base in the middle of the target sequences. Besides perfect-matched (PM) probes, MM probes are incorporated to identify any nonspecific hybridization. These tiling probes determine homozygous or heterozygous genotypes at the SNP position. The amount of information can be limited compared with the resequencing microarrays, as genotyping arrays only detect specific allele sequences that the array was designed to detect. Therefore any alleles present in a given sample, which are not included in the microarray, will not be detected [40]. However, the genotyping microarrays with sliding probe sets provide greater specificity than the resequencing microarray. In addition, recently introduced genotyping microarrays are found to include only PM probes without the incorporation of mismatch probes or shifted probes (Fig. 13.9C). Thus the number of probes per SNP in genotyping microarrays is much less than that in the resequencing microarrays. This allows for a more cost-effective addition of more SNP probes to a single genotyping microarray at the cost of reduced specificity.

Both aCGH and SNP microarrays have advantages and disadvantages. SNP microarrays were originally designed to determine simultaneously a number of genotypes but have also been utilized to provide CNV information. The signal intensity differences at each SNP probe are compared between microarrays or within a microarray, and the differences are interpreted as CNV. Thus SNP probes provide allele frequencies along with CNV, which are widely used for cancer diagnosis. One misconception is that SNP probes are good enough for whole genome variant analyses. However, locations of SNP probes are restricted to a certain genomic region, resulting in coverage issues. In contrast, copy number probes can be designed to any genomic region but generate only CNV information without allele frequencies. Therefore SNP probes and copy number probes are incorporated into a single microarray to provide much higher resolution for DNA variants. The Affymetrix CytoScan HD microarray has 1,953,246 nonpolymorphic copy number probes and 743,304 SNP probes [41].

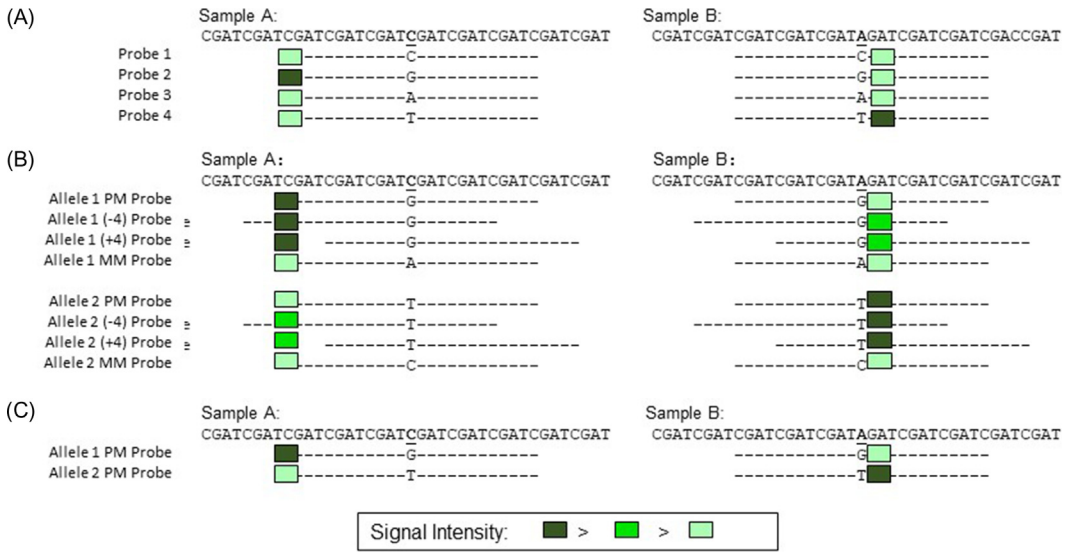


FIGURE 13.9 Microarray probe design for single nucleotide polymorphisms genotyping. (A) Resequencing microarray. (B, C) Genotyping microarray. In resequencing microarrays (A), four probes per single nucleotide polymorphisms are incorporated in the microarray. Probe sequences completely complementary to target sequences emit fluorescence signals, which are depicted by filled boxes. In old version genotyping microarrays (B), probe sequences are designed only from known allelic sequences. In the simplified depiction of the genotyping microarray (B), perfect-match and mismatch probes are incorporated. In addition, four base-shifted probes, either upstream (-4) or downstream (+4) from the middle position, are included. Complementary probes hybridize to their target sequences except for the mismatch probe (*upper left and lower right panels in B*). Polymorphic perfect-match and mismatch probes did not hybridize (*upper right and lower left panels*). But shifted probes at least hybridize to the target sequences with greater intensity than perfect-match and mismatch probes that are shown with gray boxes. That is because the shifted polymorphic site from the middle of the probe allows weaker hybridization. In recently introduced high-density genotyping microarrays as in (C), only perfect-match probes are incorporated.

Sequencing

While qPCR and dPCR assays are highly sensitive and specific, they require a prior knowledge of the target sequence. In contrast, sequencing methods are an unbiased approach to nucleic acid detection. Sequencing is widely used in clinical diagnostic laboratories to allow for identification of nucleic acid changes in genes, determine associations with diseases and phenotypes, and also identify appropriate drug targets. There are a number of different approaches to sequencing and analysis, and the next section will provide a detailed overview of the methodologies currently used in the clinical lab.

Sanger sequencing

Sanger sequencing was first developed in the 1970s and was the primary method of sequencing up until the advent of NGS in the early 2000s. The Sanger method utilizes the natural properties of DNA polymerase as well as terminator dideoxynucleotides (ddNTP) tagged with specific fluorescent dyes to generate sequence data using capillary-array electrophoresis [42]. The Sanger method makes use of the 2′,3′-dideoxy and arabinonucleoside analogs of the normal deoxynucleoside triphosphates, which act as specific chain-terminating inhibitors of DNA polymerase by preventing phosphodiester bond creation due to the absence of the 3′-hydroxyl group.

Sanger sequencing starts with amplification and purification of the target DNA. A primer is then annealed adjacent to the sequence of interest and extended by DNA polymerase. During the extension reaction, the nascent chain is terminated by the random incorporation of fluorescently labeled ddNTP, which are complementary to the base on the opposite strand. The different sized products are then resolved by capillary-array electrophoresis, and the resultant pattern of fluorescent peaks determines the DNA sequence. The technique is rapid, robust, accurate, and can typically achieve read lengths of up to 1 kb with relatively low cost. The entire Human Genome Project was completed using Sanger sequencing, and it remains the "gold standard" of DNA sequence analysis [1].

Next-generation sequencing

The complexity and duration (~10 years) of the Human Genome Project demonstrated the need for higher throughput sequencing technologies that were capable of sequencing more than one target at a time in a single run. NGS was being made available just as the Human Genome Project was wrapping up. This technology, also known as the second-generation sequencing, allowed scientists to

sequence an entire genome in one run through the use of massively parallel sequencing of millions of DNA templates. NGS rapidly produces an enormous volume of data at a lower cost than Sanger sequencing.

The commercially available NGS platforms differ from traditional Sanger sequencing technology in a number of ways. First, the DNA sequencing libraries are clonally amplified in vitro, eliminating the need to clone the DNA library into bacteria [43]. Second, the DNA is sequenced by synthesis (addition of complementary nucleotides), rather than through chain termination chemistry. Finally, the use of molecular barcodes allows the DNA templates to be sequenced simultaneously in a parallel fashion.

The workflow for NGS consists of four main steps: library preparation, template amplification, sequencing, and data analysis (Fig. 13.10). Extensive reviews have been published on the various NGS platforms [44–46]. Due to the rapidly evolving technologies, we encourage readers to visit the website of the commercially available platforms for the latest information (Ion Torrent, Illumina, Nanopore, and PacBio), and we will focus on the basic principles used for sequencing.

Library preparation

DNA library preparation usually involves fragmentation of isolated DNA into smaller, random, overlapping fragments followed by end repair and adapter ligation. Depending on the platform used, DNA is fragmented usually by physical/mechanical (i.e., ultrasonication) or enzymatic methods down to the range of 150–800 bp. While direct sequencing of RNA molecules is possible, many transcriptome-based NGS protocols will use reverse transcription to convert RNA into cDNA for sequencing [47]. cDNA libraries are prepared by removing rRNA, synthesizing cDNA, end polishing cDNA fragments (blunt-end conversion), and attaching platform-specific adapter sequences. Since the average size of RNA is smaller than DNA, no fragmentation is usually required. The cDNA library preparation method varies depending on the RNA species under investigation, which can differ in size, sequence, structural features, and abundance. Custom gene panels can be performed using libraries that are enriched with targeted gene sequences.

Amplification-based enrichment methods

Amplicon-based NGS methods use PCR to generate the sequencing libraries and are ideal for sequencing only specific ROIs (Fig. 13.11A). They are ideal when analyzing archived tissues that have limited material and rely on carefully designed sets of primers to avoid or minimize the amplification of pseudogenes (or genomic regions with high sequence homology to the ROI). At least four amplicon-based library preparation approaches are currently being used for preparation and enrichment of target sequences of interest: multiplex PCR, single-plex PCR, targeted capture followed by multiplex PCR, and anchored multiplex PCR [46,48]. Amplicon-based enrichment strategies can achieve 100% coverage of a region with little off-target sequence as the start and stop coordinates of each amplicon are predetermined. However, one limitation of amplification-based target enrichment is its susceptibility to allele dropout, which can result in false-negative calls. Allele dropout occurs when the variant is located in the primer-binding site, leading to failed amplification and allele bias [49].

Hybrid capture enrichment methods

Specific nucleotide probes are designed to pull out or capture specific ROIs by hybridization (Fig. 13.11B). These probes are much longer than typical PCR primers and, therefore, are not as affected by variants in the probe-binding site, avoiding dropout issues. The disadvantage is that probes can tolerate mismatches; thus there is less specificity, increasing off-target reads. Hybrid capture enrichment also requires more starting material than amplification-based enrichment, which could pose a problem if the amount of clinical sample is limited.

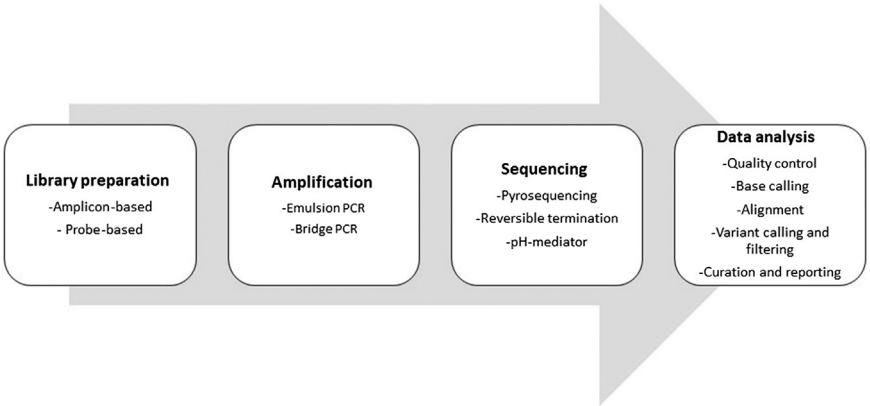


FIGURE 13.10 Overview of next-generation sequencing.

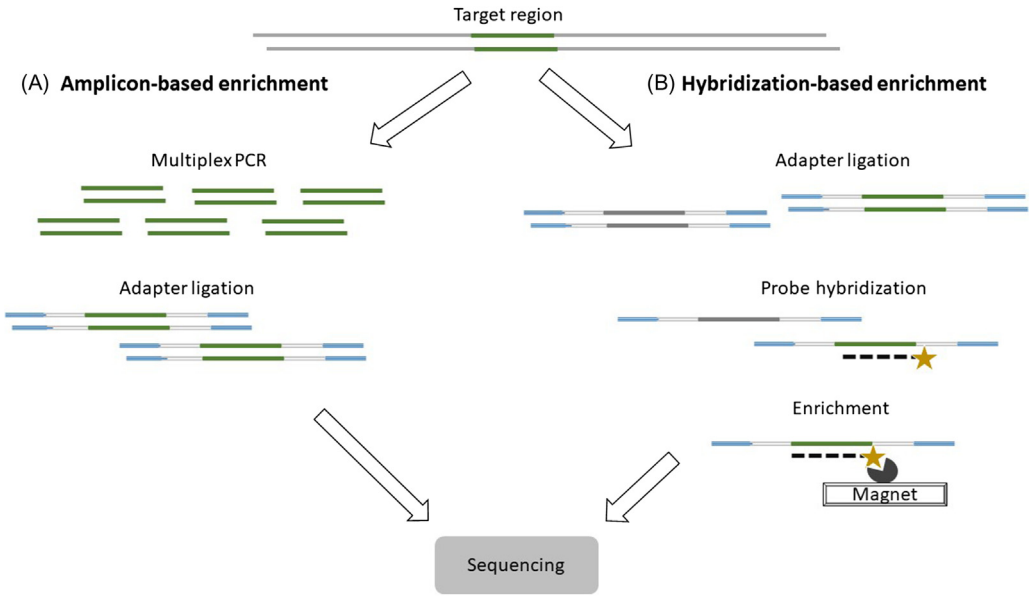


FIGURE 13.11 Schematic representation of the differences between (A) amplicon-based library preparation and (B) hybridization-based library preparation.

Template amplification

Template amplification allows for a significant increase in the available molecules for sequencing. DNA fragments selected after library preparation or specific amplification are subjected to clonal amplification. Clonal amplification involves solid phase amplification of the template and allows a strong detectable signal during sequencing. Single DNA fragments to be sequenced are either bound to beads or flow cells. Depending on the sequencing platform, emulsion PCR (emPCR) or bridge PCR is used to amplify the anchored DNA fragments into millions of spatially separated template fragments.

Emulsion polymerase chain reaction

emPCR is an alternative method that some NGS technologies utilize to replicate targeted DNA sequences. The adapter sequences ligated onto the ends of the DNA template hybridize to capture oligonucleotides covalently linked to beads. The template molecules and beads are then mixed to achieve an average of one template molecule per bead in an individual compartment created by the emulsion in an oil and water mixture. During emPCR, the surface of the bead becomes coated with clonal copies of the DNA template.

Isothermal bridge amplification (bridge polymerase chain reaction)

In isothermal bridge amplification, a method utilized by Illumina, the DNA template is first denatured, and then, during sequencing, the adapter sequences hybridize to complementary capture oligonucleotides covalently linked to the surface of the flow cell. These captured oligonucleotides are used as primers for the PCR amplification. After a second denaturation, the newly synthesized strand can then bend to hybridize with an adjacent capture oligonucleotide primer, and the cycle can be repeated to generate a cluster of identical template molecules [45]. Bridge PCR allows paired-end sequencing (sequence the DNA fragment from both ends) resulting in high coverage, a high number of reads, and more data as compared with single-end sequencing systems. This approach allows for the detection of genomic rearrangements (insertions, deletions, and inversions), repetitive sequence elements, gene fusions, and novel transcripts [50]. The majority of the sequencing chemistries use sequencing by synthesis.

Sequencing by reversible termination or sequencing by synthesis

This technique uses reversible, fluorescently labeled terminator nucleotides that are modified at the 3′ end with a cleavable terminator moiety to ensure that only a single nucleotide incorporation event can occur with each sequencing cycle [50]. Sequencing by reversible termination starts with the annealing of sequencing primer to the template molecule. Following this annealing event, each cycle consists of three steps: (1) incorporation of the complementary reversible terminator nucleotide by DNA polymerase into the DNA strand; (2) detection of the fluorescence signal for each cluster; and (3) the cleavage and removal of the fluorescent label and 3′ terminator moiety,

leading to the regeneration of the growing strand for a subsequent cycle. Repetition of these cycles leads to sequencing of the template. Natural competition between all four of the nucleotides present during each sequencing cycle reduces the inherent bias.

pH-mediated sequencing

This method is based on the detection of hydrogen ions liberated upon incorporation of each nucleotide and is not dependent on altered nucleotides, enzymes, or optical detection. Template DNA obtained after library preparation and clonal amplification is bound to ion sphere particles. Ideally, each bounded sphere is separated in a unique microwell of a chip. Then, a single type of dNTP is flowed on the chip at a time. If there is incorporation of the specific dNTP during DNA synthesis, the release of a hydrogen ion induces a change in pH, detected by ion-sensitive field-effect transistor sensors, which translates the chemical signal into digital signal, measured within seconds [51]. This technique is rapid as sequencing is done in real time with read length up to 400 nucleotides and operates at low cost, since it does not incorporate modified bases [52]. In this sequencing chemistry, homopolymer sequencing is error-prone, as repeats will be incorporated in one cycle, which leads to a proportionally higher electronic signal due to the corresponding number of released hydrogen ions. ThermoFisher has come up with the Hi-Q polymerase that has improved sequencing accuracy, decreased insertion–deletion (indel) error rates, and reduced GC-bias.

Data analysis

Data analysis consists of three steps: base calling and quality score computation; assembly and alignment; and variant calling and annotation [53]. Each NGS platform has a unique data processing pipeline, using their own proprietary analysis software to call base pairs and generate associated quality scores. The data output from an NGS instrument essentially consists of a text file of millions of raw sequence reads. The massive amount of sequence data generated by NGS creates new challenges for the laboratory, requiring significant investment in bioinformatics infrastructure and personnel. The accurate interpretation of the data generated by NGS is labor intensive, requiring expertise in genetics, pathology, clinical medicine, and informatics. The establishment of well-curated genomic databases with phenotypic information is a critical step for the widespread clinical implementation of NGS-based tests [54].

Third-generation sequencing

While NGS has revolutionized biomedical research and is now being used routinely in the clinical diagnostic space, the "third" generation of sequencing was introduced to the market in 2015 and is likely to have a role in the clinical lab in the near future. Developed by Oxford Nanopore Technologies Limited, the "MinION" is different from other sequencing technologies, because it utilizes nanopores for sequencing [55]. Able to fit in the palm of your hand and connected to a laptop computer with a USB connection, the MinION allows for direct, electronic analysis of single molecules in real time. It can be used for the analysis of DNA, RNA, proteins, and small molecules with a range of applications in personalized medicine, agriculture, and scientific research.

Implementation of molecular assays in the clinical laboratory

There is a wide variety of molecular assays that can be implemented in the clinical laboratory. In addition, the rapid evolution of sequencing and other molecule-based technologies is constantly changing the field. While the number of molecular diagnostic tests being performed for specific applications continues to increase, the number of FDA-approved or cleared tests represents only a small percent of those molecular tests conducted in laboratories. Most high volume tests that could financially justify submission to the FDA based on test volumes have been cleared and mainly affect infectious disease testing. The vast majority of molecular diagnostic tests fall into the LDT regulatory category and require significant analytical validation by the laboratory before they are used in patient care. There are numerous guidelines for validation of such LDTs; the College of American Pathologists has checklists specific for the different aspects of molecular testing. While there are current procedural terminology codes for various types of molecular tests based on technology and gene or variant tested, often payors use the "no FDA-approval" rationale to not reimburse laboratories for this test. This short-sightedness has a significant impact on patient care, as most of these tests have proven clinical utility and nonpayment limits access to these tests for which no FDA-approved options exist. For more common molecular diagnostic tests, reimbursement rates may cover the cost of performing the test, while for rare gene tests or more complex sequencing and array-based tests, reimbursement rates, if any are paid, are not even close to covering the costs.

Conclusion

Clinical molecular diagnostics is a rapidly growing and continuously changing field. The past 10 years in particular has seen significant improvements in the speed in which laboratory scientists can study the human genome. As discussed in this chapter, there are a wide variety of

nucleic-acid-based tests performed in the clinical lab, which are used to diagnose and monitor disease, detect risk, and decide which therapies will work best for individual patients. From PCR-based genotyping to the latest advances in whole genome sequencing, our knowledge of how specific biological markers play a role in disease has greatly contributed to precision medicine.

References

[1] C. Venter, M. Adams, E. Myers, P. Li, R. Mural, G.G. Sutton, et al., The sequence of the human genome, Science 291 (2001) 1304–1351.

[2] Consortium THMP, Structure, function and diversity of the healthy human microbiome, Nature 486 (2013) 207–214.

[3] J. Watson, F. Crick, Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid, Nature 171 (1953) 737–738.

[4] F. Crick, J. Watson, The complementary structure of deoxyribonucleic acid, Proc. Natl Acad. Sci. 223 (1954) 80–96.

[5] E. Southern, Detection of specific sequences among DNA fragments separated by gel electrophoresis, J. Mol. Biol. 98 (1975) 503–517.

[6] J. Stacey, P. Isaac, Restriction enzyme digestion, gel electrophoresis, and vacuum blotting of DNA to nylon membranes, Methods Mol. Biol. 28 (1994) 25–36.

[7] P.R. Langer-Safer, M. Levine, D.C. Ward, Immunological method for mapping genes on Drosophila polytene chromosomes, Proc. Natl Acad. Sci. 79 (1982) 4381–4385.

[8] R.K. Saiki, D.H. Gelfand, S. Stoffel, S.J. Scharf, R. Higuchi, G.T. Horn, et al., Primer directed enzymatic amplification of DNA with a thermostable DNA polymerase, Science 239 (1988) 487–491.

[9] R.K. Saiki, S. Scharf, F. Faloona, K.B. Mullis, G.T. Horn, H.A. Erlich, et al., Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia, Science 230 (1985) 1350–1354.

[10] K. Mullis, F. Faloona, Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction, Methods Enzymol. 155 (1987) 335–350.

[11] C. Vieille, G.J. Zeikus, Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability, Microbiol. Mol. Biol. Rev. 65 (2001) 1–43.

[12] K. Terpe, Overview of thermostable DNA polymerases for classical PCR applications: from molecular and biochemical fundamentals to commercial systems, Appl. Microbiol. Biotechnol. 97 (2013) 10243–10254.

[13] R.B. Wallace, J. Shaffer, R.F. Murphy, J. Bonner, T. Hirose, K. Itakura, Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effects of single base pair mismatch, Nucl. Acids Res. 6 (1979) 3543–3558.

[14] P. McInerney, P. Adams, M.Z. Hadi, Error rate comparison during polymerase chain reaction by DNA polymerase, Mol. Biol. Int. 2014 (2014) 287430.

[15] M.J. Wacker, M.P. Godard, Analysis of one-step and two-step real-time RT-PCR using SuperScript III, J. Biomol. Tech. 16 (2005) 266–271.

[16] A. Smith, R. Nelson, Capillary electrophoresis of DNA, Curr. Protoc. Mol. Biol (1999) 2.8.1–2.8.17.

[17] B.C. Durney, C.L. Crihfield, L.A. Holland, Capillary electrophoresis applied to DNA: determining and harnessing sequence and structure to advance bioanalyses (2009–2014), Anal. Bioanal. Chem. 407 (2015) 6923–6938.

[18] Y. Song, P. Roumagnac, F. Weill, J. Wain, C. Solecek, C. Mazzoni, et al., A multiplex single nucleotide polymorphism typing assay for detecting mutations that result in decreased fluoroquinolone susceptibility in *Salmonella enterica* serovars *Typhi* and *Paratyphi*, J. Antimicrob. Chemother. 65 (2010) 1631–1641.

[19] J. Wilhelm, A. Pingoud, Real-time polymerase chain reaction, Chem. Biochem. 4 (2003) 1120–1128.

[20] C. Wittwer, M. Herrmann, A. Moss, R. Rasmussen, Continuous fluorescence monitoring of rapid cycle DNA amplification, Biotechniques 22 (1997) 130–138.

[21] E. Navarro, G. Serrano-Heras, M. Castano, J. Solera, Real-time PCR detection chemistry, Clin. Chim. Acta 439 (2015) 231–250.

[22] D. Klein, Quantification using real-time PCR technology: applications and limitations, Trends Mol. Med. 6 (2002) 257–260.

[23] R.T. Ranasinghe, T. Brown, Fluorescence based strategies for genetic analysis, Chem. Commun. (2005) 5487–5502.

[24] D.J. French, C.L. Archard, T. Brown, D.G. McDowell, HyBeacon™ probes: a new tool for DNA sequence detection and allele discrimination, Mol. Cell. Probes 15 (2001) 363–374.

[25] C. Wittwer, G. Reed, C. Gundry, J. Vandersteen, R. Pryor, High resolution genotyping by amplicon melting analysis using LC Green, Clin. Chem. 49 (2003) 853–860.

[26] M. Erali, K.V. Voelkerding, C.T. Wittwer, High resolution melting applications for clinical laboratory medicine, Exp. Mol. Path 85 (2008) 50–58.

[27] G. Geiss, R. Bumgarner, B. Birditt, T. Dahl, N. Dowidar, D. Dunaway, et al., Direct multiplexed measurement of gene expression with color-coded probe pairs, Nat. Biotech. 26 (2008) 317–325.

[28] E. Mokany, S.M. Bone, P.E. Young, T.B. Doan, A.V. Todd, MNAzymes, a versatile new class of nucleic acid enzymes that can function as biosensors and molecular switches, J. Am. Chem. Soc. 132 (2010) 1051–1059.

[29] M. Al-Turkmani, K. Godwin, J. Peterson, G. Tsongalis, Rapid somatic mutation testing in colorectal cancer by use of a fully automated system and single-use cartridge: a comparison with next-generation sequencing, J. Appl. Lab. Med. 3 (2018) 1–7.

[30] S.A. Turner, S.J. Deharvengt, K.D. Lyons, J.A.P. Espinal, E.P.M. LaRochelle, S. Bejarano, et al., Implementation of multicolor melt curve analysis for high-risk human papilloma virus detection in low- and middle-income countries: a pilot study for expanded cervical cancer screening in Honduras, J. Glob. Oncol. 4 (2017) 1–8.

[31] P. Sykes, S. Neoh, M. Brisco, E. Hughes, J. Condon, A. Morley, Quantitation of targets for PCR by use of limiting dilution, Biotechniques 13 (1992) 444–449.

[32] L. Cao, X. Cui, J. Hu, Z. Li, J.R. Choi, Q. Yang, et al., Advances in digital polymerase chain reaction (dPCR) and its emerging biomedical applications, Biosens. Bioelect 90 (2017) 459–474.

[33] A.S. Whale, J.F. Huggett, S. Tzonev, Fundamentals of multiplexing with digital PCR, Biomol. Detect. Quant. 10 (2016) 15–23.

[34] D. Gresham, M.J. Dunham, D. Botstein, Comparing whole genomes using DNA microarrays, Nat. Rev. Genet. 9 (2008) 291–302.

[35] N.P. Carter, Methods and strategies for analyzing copy number variation using DNA microarrays, Nat. Genet. 39 (2007) 16–21.

[36] A. Kallioniemi, O. Kallioniemi, D. Sudar, D. Rutovitz, J. Gray, F. Waldman, et al., Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors, Science 258 (1992) 818–821.

[37] A.A. Dombkowski, B.J. Thibodeau, S.L. Starcevic, R.F. Novak, Gene-specific dye bias in microarray reference designs, FEBS Lett. 560 (2004) 120–124.

[38] C. Reilly, A. Raghavan, P. Bohjanen, Global assessment of cross-hybridization for oligonucleotide arrays, J. Biomol. Tech. 17 (2006) 163–172.

[39] J.G. Hacia, J.B. Fan, O. Ryder, L. Jin, K. Edgemon, G. Ghandour, et al., Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays, Nat. Genet. 22 (1999) 164–167.

[40] R. Bumgarner, DNA microarrays: types, applications and their future, Curr. Protoc. Mol. Biol. 101 (2013) 22.1.1–22.1.11.

[41] H.-S. Jung, J.A. Lefferts, G.J. Tsongalis, Utilization of the oncoscan microarray assay in cancer diagnostics, Appl. Cancer Res. 37 (2017) 1.

[42] F. Sanger, A.R. Coulson, A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase, J. Mol. Biol. 94 (1975) 441–448.

[43] M.W. Anderson, I. Schrijver, Next generation DNA sequencing and the future of genomic medicine, Genes 1 (2010) 38–69.

[44] J. Xuan, Y. Yu, T. Qing, L. Guo, L. Shi, Next-generation sequencing in the clinic: promises and challenges, Cancer Lett. 340 (2013) 284–295.

[45] H.P.J. Buermans, J.T. den Dunnen, Next generation sequencing technology: advances and applications, Biochim. Biophys. Acta 1842 (2014) 1932–1941.

[46] F. Chang, M.M. Li, Clinical application of amplicon-based next-generation sequencing in cancer, Cancer Gen. 206 (2013) 413–419.

[47] R. Hrdlikova, M. Toloue, B. Tian, RNA-Seq methods for transcriptome analysis, Wiley Interdiscip. Rev. RNA 8 (2017) 1–12.

[48] Z. Zheng, M. Liebers, B. Zhelyazkova, Y. Cao, D. Panditi, K.D. Lynch, et al., Anchored multiplex PCR for targeted next-generation sequencing, Nat. Med. 20 (2014) 1479–1484.

[49] S. Hahn, A. Garvin, D. Naro, W. Holzgreve, Allele drop-out can occur in alleles differing by a single nucleotide and is not alleviated by preamplification or minor template increments, Genet. Test. 2 (1998) 351–355.

[50] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, et al., Accurate whole human genome sequencing using reversible terminator chemistry, Nature 456 (2008) 53–59.

[51] J.M. Rothberg, W. Hinz, T.M. Rearick, J. Schultz, W. Mileski, M. Davey, et al., An integrated semiconductor device enabling non-optical genome sequencing, Nature 475 (2011) 348–352.

[52] S. Ambardar, R. Gupta, D. Trakroo, R. Lal, J. Vakhlu, High throughput sequencing: an overview of sequencing chemistry, Indian J. Microbiol. 56 (2016) 394–404.

[53] S.P. Pfeifer, From next-generation resequencing reads to a high-quality variant data set, Heredity 118 (2016) 111–124.

[54] R. Cook-Deegan, J.M. Conley, J.P. Evans, D. Vorhaus, The next controversy in genetic testing: clinical data as trade secrets? Eur. J. Hum. Genet. 21 (2013) 585–588.

[55] T. Laver, J. Harrison, P.A. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz, et al., Assessing the performance of the Oxford Nanopore Technologies MinION, Biomol. Detect. Quant. 3 (2015) 1–8.

Self-assessment questions

1. A nucleotide is composed of all of the following except:
   a. phosphate group
   b. pentose sugar
   c. base
   d. fluorophore
2. The discreet structural units of the human genome are known as:
   a. nucleotides
   b. chromosomes
   c. bases
   d. Z DNA
3. Hybridization of probes to target sequences is possible due to:
   a. the presence of coding sequences
   b. the presence of noncoding sequences
   c. complementary base pairing
   d. all of the above
4. Electrophoresis of DNA for a Southern blot assay is accomplished by first:
   a. amplifying target sequences
   b. digesting DNA with restriction enzymes
   c. transferring the DNA to a membrane
   d. UV crosslinking
5. Primers used for the process of PCR are.
   a. single-stranded DNA oligonucleotides
   b. dsDNA oligonucleotides
   c. Single-stranded RNA oligonucleotides
   d. double-stranded RNA oligonucleotides
6. Detection systems used in real-time PCR are based on:
   a. radioactivity
   b. chemiluminescence
   c. fluorescence
   d. colorimetry
7. Which type of real-time PCR probe is dependent on exonuclease activity of the polymerase:
   a. Scorpion probes
   b. TaqMan probes
   c. FRET probes
   d. molecular beacons
8. Which of the following is NOT true about ddPCR?
   a. It quantifies the nucleic acid target of interest
   b. It is based on amplification of a single copy of DNA in an oil droplet
   c. DNA must first be digested with restriction enzymes
   d. It eliminates primer bias
9. Which two probes are incorporated in microarrays for genomic aberrations?
   a. copy number probes
   b. SNP probes
   c. exon probes
   d. gene probes
10. Next-generation sequencing set up requires:
    a. library preparation in the wet lab
    b. extensive understanding in bioinformatics
    c. multiple quality control steps during both library prep and bioinformatics
    d. all of the above

Answers

1. d
2. b
3. c
4. b
5. a
6. c
7. b
8. c
9. a and b
10. d

Chapter 14

Laboratory automation

Mark A. Marzinke

Departments of Pathology and Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Define preanalytical, analytical, and postanalytical tasks that may be automated.
- Distinguish between task-targeted, subtotal, and total laboratory automation.
- Identify the components of centralized laboratory automation systems.
- Explain the expectations and considerations associated with the implementation of integrated automation into the clinical laboratory.
- Discuss methods and approaches to monitor integrated automation performance and success postimplementation.

The global landscape of healthcare is in a perpetual state of transformation, and the clinical laboratory must adapt and evolve accordingly. Healthcare reforms, modifications to reimbursement plans, staffing shortages with increased demands, and fiscal constraints are just some of the factors that require clinical laboratories to subscribe to the "do more with less" ideology. One way in which laboratories have responded to such changes is via the adoption and incorporation of laboratory automation. While automation was once considered a luxury for larger clinical or reference laboratories, its use, in various iterations, has become pervasive in today's clinical environment. Currently, it would be rare to find a laboratory setting that did not rely on some form of automation, either during the analytical phase of testing or applied to pre- or postanalytical processes. Further, current estimates indicate that more than 30% of clinical laboratories in North America, Asia, and Europe have incorporated some degree of centralized automation into their laboratory workflows.

The transition from manual techniques to an automated platform is not new. For the clinical chemist, the beginnings of laboratory automation can be traced back to the mid-1950s with the introduction of AutoAnalyzer I, an analytical platform that utilized the discrete separation of a continuous stream of materials via air bubbles to perform sequential sample testing. AutoAnalyzer I exploited the principle of continuous flow analysis (CFA) for increased analytical throughput; thus the ability to automate the analytical phase of the testing process was readily adopted by clinical laboratorians for the measurement of routine chemistry analytes. While several iterations of automated analyzers followed in the coming decades, the field pivoted in the 1980s due to advances in centralized automation by Dr. Masahide Sasaki and colleagues. Using a combination of conveyor belts, basic circuitry, and robotics, specimens could be loaded and unloaded onto clinical analyzers. In addition, these activities were directed by process control software, also known as the laboratory automation software. Dr. Sasaki transitioned automation beyond the analytical testing phase; preanalytical automation and postanalytical automation were now possible and successful in a clinical laboratory setting. Three decades following the incorporation of "home-brewed" preanalytical automation and postanalytical automation within Dr. Sasaki's laboratory, all major vendors now offer fully integrated automation solutions for clinical laboratory testing.

The complementarity of work performed between the analytical phase and the pre- and postanalytical phases of the testing process led to the advent of total laboratory automation (TLA), an all-inclusive, assembly line-like approach toward laboratory activities. Currently, most automated analyzers are configured to perform random-access sample analysis, in which each specimen is independently analyzed for various test combinations. Automated analyzers can accommodate multiple analyte-specific reagents to facilitate the multiplexed testing of more than one analyte within a specimen. This is in contrast to batch-type analysis, in which specimens are grouped for testing; batch analysis can affect throughput and prolong result turnaround time (TAT).

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00014-4

Goals of automation

In many ways, laboratory automation is the ideal demonstration of the Lean Six Sigma philosophy, where efficiency is maximized and waste and errors are minimized. This is achieved through the integration of robotics, circuitry, and informatics to perform routine laboratory tasks in a standardized manner, thus circumventing the intra- and interindividual variabilities associated with manual processes. There are numerous pre- and postanalytical processes that are ripe for automation; these are summarized in Table 14.1.

Automation may be presented to the clinical laboratory in several ways; however, the utilization and incorporation of integrated instrumentation must be in line with the goals and activities of the laboratory. Globally, automation can be stratified as task-targeted, subtotal, or total in nature, depending on the activities and processes supported by the platform. Task-targeted automation, as the name implies, encompasses the automation of specific laboratory activities, for example, a task-targeted instrument may focus solely on the sorting of specimens for downstream laboratory processes. Subtotal or stand-alone platforms involve the automation of a number of tasks in order to streamline the overall laboratory workflow. Subtotal automation is not physically connected to the laboratory analyzer by a track system; thus the entire testing process is not automated, as specimens must still be transported manually to the testing area. TLA platforms offer all of the activities contained within subtotal automation, with the additional feature of physical connectivity with the laboratory analyzer via a track system. Centralized automation systems may be dichotomized as either open or closed systems. Open systems offer software and hardware flexibilities, so instruments from other vendors may be incorporated into the automated track system, or "line." Closed systems do not offer this flexibility, and all of the process control software and preanalytical, analytical, and postanalytical instrumentation are controlled by a single vendor. As of 2018, all major laboratory vendors, including Abbott Diagnostics, Beckman Coulter, Roche Diagnostics, and Siemens Healthineers, offer centralized automation solutions to optimize laboratory workflows. Further, many vendors also demonstrate open access and flexibility to facilitate the integration of third-party systems onto their automated lines [1]. This provides a notable advantage in the consolidation of testing across laboratory disciplines and subdisciplines.

Although the specific requirements for centralized laboratory automation may differ based on the specific clinical laboratory environment, many of the evaluable criteria are universal. Common considerations include the electronic identification and recognition of specimens and orders, specimen container flexibility, prioritization of sample management, and both physical and software-mediated integrations with analytical instrumentation, among others. These considerations will be discussed later in this chapter. Several reports, including those from Melanson and colleagues, describe processes for the evaluation and selection of central automation platforms [2]. While the capacity and functionality of each system differ among vendors, this chapter will discuss the general aspects of laboratory automation, as well as considerations and limitations with regard to introducing automation into a clinical laboratory setting.

Components of laboratory automation

There are a number of physical considerations and components that are associated with TLA systems. One prerequisite for centralized automation is the integrated capability for TLA systems to identify specimen-specific

TABLE 14.1 Preanalytical, analytical, and postanalytical laboratory activities that may be automated.

Preanalytical phase	Analytical phase	Postanalytical
Specimen receipt	Analyzer loading	Result reporting
Specimen accessioning	Specimen integrity checks	Recapping
Labeling	Testing	Specimen archiving
Sorting	Dilutional analysis	
Centrifugation	Result transmission	
Decapping	Specimen unloading	
Specimen quality inspection		
Aliquoting		
Specimen distribution		
Specimen retrieval		

characteristics from a barcoded sample. This is achieved through an interface between the process control software within the integrated automation platform and the laboratory information system (LIS). As barcode readers scan specimens placed on an automated platform, characteristics from the LIS, including order-specific information such as specimen container type, tests requested, required testing volumes, and the number of aliquots required by the analyzing instrument, will be used by the automated platform to direct follow-up activities. Of note, the process control software must be flexible enough to mask a module or portion of the automated platform while continuing other laboratory tasks. This flexibility facilitates minimally interrupted workflow if a specific portion of the centralized system requires maintenance, service, or other troubleshooting.

Because specimen barcodes direct the downstream activities of centralized laboratory automation, nonbarcoded specimens are suboptimal candidates for automation. Specimens received in this manner require additional manual processing, including the accessioning of pertinent information into the LIS as well as labeling of accessioned specimens by laboratory staff. A general schematic of the various stations or modules that are typically included in centralized laboratory automation solutions is illustrated in Fig. 14.1. However, it is noted that each vendor has its own specific performance footprint, characteristics, and workflows. An overview of commonly included stations will be discussed in more detail in this section:

1. ***Specimen loading area:*** Upon receipt in the laboratory, barcoded specimens must be placed in a transport carrier to facilitate subsequent processing steps. Nonbarcoded specimens may be accessioned into the LIS for the generation of an order (or specimen) label for instrument recognition. Centralized automation transport carriers vary by vendor and can range from single specimen carriers, which accommodate one specimen at a time, to racks that may accommodate 5−10 specimens. Every TLA platform requires (manual or automated) specimen loading into a carrier to facilitate subsequent activities on the system and movement within the track system. To stratify and triage testing priority between routine and stat, or emergent, samples, most automation platforms also offer two input areas to prioritize the processing of stat samples.
2. ***Barcode readers:*** Barcode readers are associated with the specimen loading area so that the aforementioned specimen-specific information may be directed from the LIS to the process control software of the automation platform. Barcode scanning by the TLA instrumentation also serves as an electronic receipt within the laboratory and can be used in the determination of laboratory receipt to result TAT or may be helpful in workflow-driven investigations. Platforms may have multiple barcode readers to track and direct downstream processes or a single barcode reader, in which the sample is scanned once in the specimen loading module. In the latter case, specimen-specific information may become associated with its location within the transport container; alternatively, the information may be linked via a radiofrequency ID chip. The standard and preferred barcode symbology used in integrated automation systems is Code 128, which is a high-density symbology with the capacity to encode all 128 characters

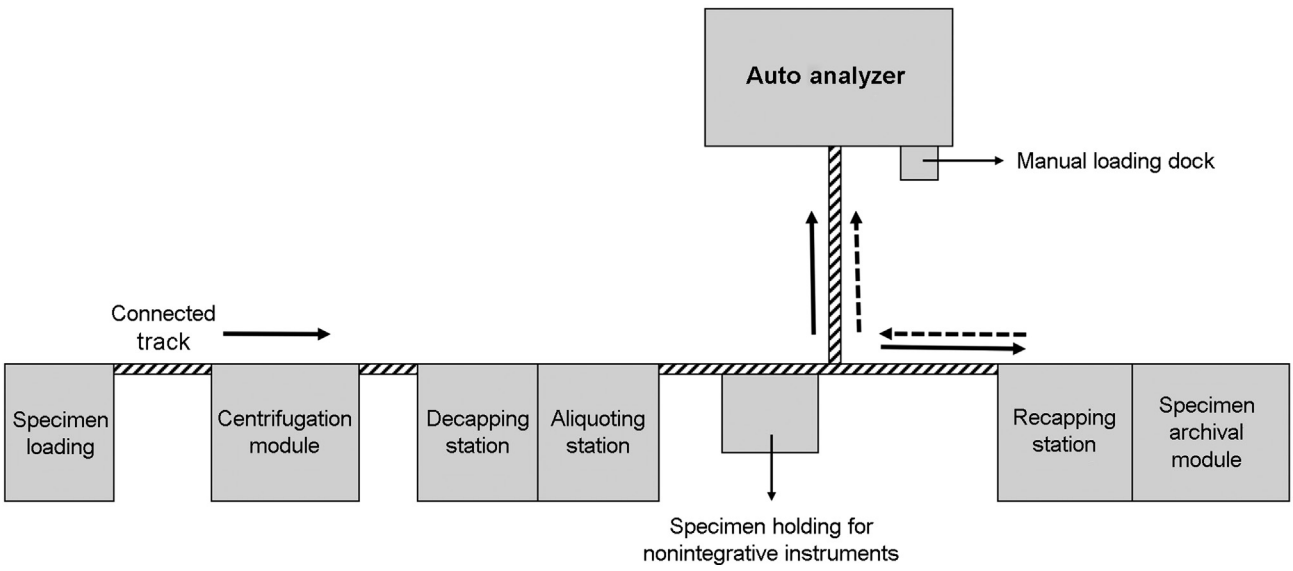


FIGURE 14.1 Representative layout of a centralized automation platform. Specific modules and functionality are vendor-specific.

from the American Standard Code for Information Interchange [3]. Several TLA platforms and LIS systems now recognize 2D barcodes, which include higher density-encoded information; however, specific barcode scanners are required for the reading of either linear (1D) or 2D barcoded samples.

3. ***Transport and sorting tracks:*** Using a conveyor-based track system, specimens are directed to other areas of instrumentation. Many TLA platforms are flexible, so that several destinations may be possible. Follow-up activities are driven by the process control software and the electronic information associated with the specimen or transport carrier. Some vendors use prebarcoded carriers or racks that will direct specimens toward specific modules within the platform. This flexibility in specimen routing facilitates the management and handling of multiple specimen types on a TLA system. For example, both specimens requiring centrifugation as well as those that do not may be processed on a TLA system. The track system must be able to accommodate both specimen types and appropriately direct the sample to a centrifuge or alternatively bypass that component of the system. Currently, TLA solutions from all major vendors have incorporated track bidirectionality into their centralized platforms, allowing for enhanced capability in automating routine laboratory workflows. Bidirectional tracks can assist in automating laboratory activities such as the fulfillment of add on requests or selecting a sample from a storage unit for retesting. The latter will be discussed later in this chapter.
4. ***Centrifugation:*** For samples that require centrifugation, specimens can be routed from the track system to the centrifuge module. Specimens in transport carriers are loaded into self-balancing centrifuges. Centrifuge throughput is platform-specific, and capacity is influenced by centrifugation speed and duration. However, the theoretical throughput of various preanalytical platforms ranges from 300 to 600 specimens per hour. A rate limiting step in preanalytical sample throughput is the volume of specimens requiring centrifugation as well as the potential test-dependent variability in centrifuge conditions. Some vendors offer TLA solutions that are modular in nature and can accommodate multiple centrifuges within a single system, thereby increasing and optimizing laboratory efficiency. However, one consideration in determining the number of centrifuge modules is the additional physical footprint and the impact additional capacity may have on space in the clinical laboratory setting. Depending on the platform, centrifugation may occur under refrigerated or room temperature conditions; most automated systems utilize the former for sample preparation. Following centrifugation, specimens are returned to the track for further preanalytical activities.
5. ***Decapping station:*** TLA decapping stations utilize robotic arms for the removal of the cap from the primary container introduced onto the platform. The cap is then disposed of in a waste repository, most commonly located within the system.
6. ***Aliquoting station:*** Following the removal of caps, specimens may be routed via the conveyor-like track system into an aliquoting station. Using robotic arms, specimen volumes are aspirated and aliquots are transferred to secondary containers. Depending on the specific layout and specifications of the system, barcode printers may be associated with this module, as secondary specimen containers may be labeled for downstream testing. Alternatively, for aliquoted samples that are tested in an integrated system and are transferred to sample cups, all of the electronic information associated with the specimen may be located within the transport carrier barcode. The process control software of the automation platform utilizes the requested test and order information to determine transfer volumes into the secondary container. Disposable tips are commonly used to prevent cross-contamination, and, postdispensing, tips are disposed of into a waste repository.

   Aliquoting apparatuses used in most current preanalytical automation platforms have the ability to detect clots; this is achieved through the sensing of increased resistance by the aliquoting probe upon aspiration. When the resistance threshold is exceeded or triggered, specimens may be defaulted from the system for investigation. When compared with manual workflows, automated aliquoting is a critical function and plays a significant role in reducing specimen-associated pour-over errors.

   All AutoAnalyzers can be sampled directly from the primary collection container, and the laboratory can opt to bypass sample aliquoting for direct downstream sampling and analysis. In the current landscape, in order to optimize efficiency and streamline intralaboratory workflows, many laboratories will directly sample from the primary container. However, if testing is performed on an automated analyzer that uses a probe-based system for sampling, carryover and contamination studies should be conducted to minimize the risk of contaminating a primary container during testing, particularly if the sample is being used for analysis in other testing areas. If aliquoting is required for offline or send out testing, this activity may be performed after integrated testing is performed on a primary container. The specific needs within the laboratory will often

drive the physical footprint and configuration of a centralized automation system, including how an aliquoting module may fit into an operational workflow.

7. ***Recapping station:*** Centralized platforms may also include a recapping module, where aliquots or primary tubes may be sealed prior to analysis. The functionality and workflow for sample recapping is dependent on the downstream activities of the specimen. For instance, if the destination of an aliquoted sample involves direct routing to an analyzer via the track system, recapping may not be necessary. However, if a primary tube is being routed to a longer-term specimen holding area or a specimen archival system, then airtight recapping is desirable. For samples that remain uncapped, stability due to exposure to ambient temperature or air must also be assessed; analytes such as CO_2 are sensitive to exposure to ambient air and may result in falsely low results. Vendors approach recapping differently, and recapping systems vary from foil-based resealing to the use of pop-top or screw-top caps; some vendors have the capacity to facilitate more than one workflow.
8. ***Specimen removal module:*** Upon the completion of specimen processing, which may involve any combination of sorting, centrifugation, decapping, or aliquoting steps along the automated track system, samples may need to be removed from the integrated platform for analysis on nontracked analyzers or for other follow-up activities, such as completing esoteric test requests. Consequently, specimens (primary tubes or aliquots) may require routing to a holding area for removal.

While the aforementioned components all encompass preanalytical laboratory processes (however, barcode reading and specimen loading areas are universal processes in all areas of the laboratory), the analytical and postanalytical areas are also included in TLA solution platforms. Common components include:

9. ***Specimen integrity checks:*** Most analytical instruments use spectrophotometric techniques to assess the quality of a specimen prior to analysis. Many chemistry and immunochemistry tests can be affected by common endogenous interfering substances, such as hemoglobin, lipids, and bilirubin. Currently, automated analytical instruments for clinical chemistry testing assess the impact of hemolysis, lipemia, or icterus on analytical assays via spectrophotometric measurements at relevant, potentially interfering, wavelengths.

In newer iterations of TLA platforms, several vendors have incorporated specimen integrity checks during the preanalytical phase of centralized automation, allowing for the identification of samples that may be hemolyzed or compromised prior to reaching an analyzer. However, the methodologies used in these check modules are commonly based on specimen opacity as opposed to measurements at specific spectral wavelengths, and therefore should be optimized based on the types of specimens handled by the laboratory. One benefit of specimen integrity evaluation in the preanalytical phase of TLA is that problem specimens can be more acutely identified and acted upon by laboratory staff.

10. ***Dilutional analysis:*** For laboratory values above the primary linearity or measuring range of an assay, dilutional analysis may be indicated as a follow-up action. Automation of this process involves the instrument or middleware-controlled dilution of a specimen with a known diluent for reanalysis, and back calculation of the analyte activity or concentration. Further, performing dilutions online can significantly reduce analytical errors and improve workflow, as samples do not have to be retrieved and further manually manipulated by staff.
11. ***Computer interfacing:*** While all systems include computer controls to execute the physical functions associated with an integrated automation system, it is imperative that there is cross-talk between the TLA platform and the LIS. Centralized laboratory automation software must be able to interface with the instrument operator as well as mainframe computers and the LIS. Standard information that is typically communicated between the instrumentation and the LIS, either directly or through middleware software programs, includes status information of each preanalytical or analytical analyzer, specimen type, patient demographics, and data related to specimen orders, tests, results, as well as any rules to control follow-up activities, including sample reruns or dilutions.

Within the last three years, the major TLA vendors have marketed and implemented more dynamic and adaptive interfaces to allow for greater control of automation systems. Notably, these interfaces can be used to extract important laboratory performance parameters, including TAT, productivity reports, and shifts in test performance through patient result monitoring (i.e., moving averages). Enhanced access to these data can help streamline and optimize laboratory workflows.

12. ***Specimen archival module:*** In fully realized centralized automated workflows, the automation of specimen storage is critical to the complete integration of laboratory processes. Specimens can be directed to long-term storage via the track system in a manner similar to the track-driven routing to instrument

analyzers. Unlike the specimen removal module, which is typically nonrefrigerated and has limited capacity in terms of sample storing capacity, specimen archival systems (or stockyards) are refrigerated and can accommodate anywhere from 3000 to 30,000 specimens, depending on the vendor and system. Via the laboratory automation software or middleware-mediated interfacing with the LIS, the location and removal of specimens is performed using robotic arms. As previously mentioned, more recent configurations of centralized laboratory automation have incorporated a bidirectional track into their workflow. Bidirectional tracks allow for the robotics-driven identification and removal of a specimen from the stockyard, followed by the placement of the sample into a transport carrier and subsequent routing to the analyzing instrument for testing. Although associated with a larger footprint, bidirectional tracks further automate the overall laboratory workflow.

Expectations of automation

If thoughtfully approached, there are a number of benefits that may be achieved through the incorporation of automation into the clinical laboratory. A study from Holman and colleagues demonstrated that following the automation of numerous preanalytical processes, including sample sorting, centrifugation, decapping, labeling, and aliquoting, there was an 80% reduction in errors involving sample sorting, aliquoting, and labeling [4]. In a 2014 report published in *Clinical Laboratory News*, it was conveyed that postinstallation of laboratory automation reduced error rates by more than 70% [5]. Further, TLA systems may offer additional benefits, including the ability to manage larger workloads, accommodate staff shortages, and reduce overall operational costs, topics to be discussed later in this chapter.

Additional expectations of integrated laboratory automation include increased throughput and improved TAT. The consolidation and centralization of several manual preanalytical steps can maximize the efficiency of the overall testing system. However, it should be noted that the throughput and TAT of automated analyzers is dependent on a number of criteria, including specimen container type, instrument type, configuration complexity, and total testing volumes. Based on these variables, which are laboratory-specific, theoretical automation throughput may not reflect real-world operations. Therefore familiarity with current ordering and testing practices, including the number of samples tested during a period of time, the average number of tests associated with each specimen, and the physical location and number of analyzers utilized in the laboratory are all important factors in tempering unrealistic expectations with regard to centralized laboratory automation performance. A list of attributes to be considered for complete centralized laboratory automation is described in Table 14.2.

The move to subtotal or total automation is typically associated with a major shift in laboratory workflow and management. Further, the selection and subsequent success of automation is dependent on setting achievable and realistic expectations of the instrumentation for one's specific laboratory environment. As indicated by Boyd and Hawker, many laboratorians use an 80% benchmark to direct automation decisions; this means that if 80% of specimen containers and handling activities can be standardized, improvements will be achieved with regard to overall labor and costs associated with specimen handling and processing [6]. However, a comprehensive knowledge of both the current processes as well as laboratory needs is critical for automation success. Failures in implementation may occur for a number of reasons, including noncompatibility and suboptimal workarounds between the manual and automated processes, the lack of familiarity with the proposed automation system, and the overall loss of flexibility with other systems or open channels. Further, vendor contributions such as inadequate technical support and unforeseen maintenance and supply costs can impact automation success [7].

The selection and decision to incorporate and implement laboratory automation into the clinical laboratory is largely based on the needs of the laboratory. As described by Hawker and colleagues [8,9], a comprehensive workflow analysis of all laboratory activities, including staffing required for all laboratory functions, can identify areas for efficiency as well as predict the impact of centralized automation in the clinical environment. A simplified, representative workflow analysis is depicted in Fig. 14.2. Not surprisingly, common expectations and goals following an assessment of existing workflows are to improve laboratory efficiency. Preanalytical goals include the following: a reduction in sorting steps required prior to testing; a reduction in the number of times each sample is handled by laboratory staff, as well as the number of staff handling each specimen; streamlining the flowpath of a specimen with regard to specimen processing and delivery; decreasing laboratory staff traffic in the management and dissemination of samples. Analytical goals include centralization of laboratory testing and reduction in the number of shared specimens across instruments, specimen integrity checks, and automated dilutional analysis. Postanalytical automation goals include integration of laboratory results to the hospital information system and storage capacity and retrieval capabilities for remnant specimens.

While many steps have been successfully automated in integrated laboratory systems, there is still variability

TABLE 14.2 Metrics to be considered for automation.

Questions to consider
Space considerations
What is the footprint of the automation platform?
Will the laboratory space accommodate the automation platform's heat and electrical requirements?
Throughput and performance considerations
Is the throughput of the system amenable to the laboratory's needs?
What is the predicted impact on reducing laboratory errors?
What is the predicted impact on laboratory TAT?
What additional workload can the system accommodate above the laboratory's current operational needs?
What is the flexibility of the automated platform with regard to specimen containers?
Will the system be easily incorporated into existing LISs?
Analytical considerations
Will the automation system accommodate the test menu?
What is analyzer and component flexibility in terms of sample loading?
Can AutoAnalyzers perform on-board dilutions?
Does the platform have the capability to perform specimen integrity checks?
Maintenance and training considerations
What is the maintenance time of the entire system, and each component?
Is there service support to accommodate all facets of the automated platform on a 24/7 basis?
What degree of training is required for each component of the automated system?
Cost considerations
What is the initial investment in the automated platform?
What is the predicted cost-savings associated with the transition to automation?
How many full-time employees (FTEs) will be needed to operate the instrumentation?

TAT, Turnaround time.

as to how specimens are delivered to the clinical laboratory. Depending on the patient population that is served, laboratories may receive samples from a variety of locations. Common delivery mechanisms include pneumatic tube systems or courier services, which typically accommodate inpatient and outpatient or off-site locations, respectively, although not exclusively. However, there are nonintegrated automated alternatives to specimen delivery, including electric track vehicles and mobile robots [6]. Electric track vehicles have a larger carrying capacity than standard carriers used in a pneumatic tube system layout. Mobile robots may be used in delivering specimens to or within the laboratory; however, for optimal efficiency, samples should be delivered in batch. Thus, while mobile carriers may alleviate potential staff shortages, depending on the layout and needs of the laboratory, they may not increase overall efficiency.

Automation beyond the chemistry analyzer

In the current landscape, centralized laboratory automation has mainly involved the integrated connectivity of the pre- and postanalytical phases of the testing process to chemistry and immunochemistry analyzers. However, integrated automation for hematology, coagulation, urinalysis, and molecular and microbiological testing is also available. For example, Sysmex Corporation and Beckman Coulter offer track-based automation for hematology and urinalysis testing, respectively. Several centralized laboratory platforms now have the capability of automated specimen handling and transport to a variety of downstream instruments. As previously discussed, open access instrumentation can facilitate more integrated laboratory automation via track extension to nonchemistry and immunochemistry analyzers manufactured by other vendors.

Many of the benefits of centralized laboratory automation expand beyond the core laboratory setting. Automation has permeated labor-intensive and manually driven laboratory settings, including bacteriology and clinical microbiology. Automated tasks in clinical bacteriology include appropriate Petri dish selection, sample inoculation, spreading of inoculum on appropriate plates, as well as the labeling and sorting of media and plates. The automation of many of these tasks also provides a way to standardize assays across laboratories; for example, variability in the streaking of liquid or nonliquid specimens on plates is reduced by spreading inoculum via magnetic beads, an approach used by the automated specimen processing analyzer offered by BD Kiestra. Other

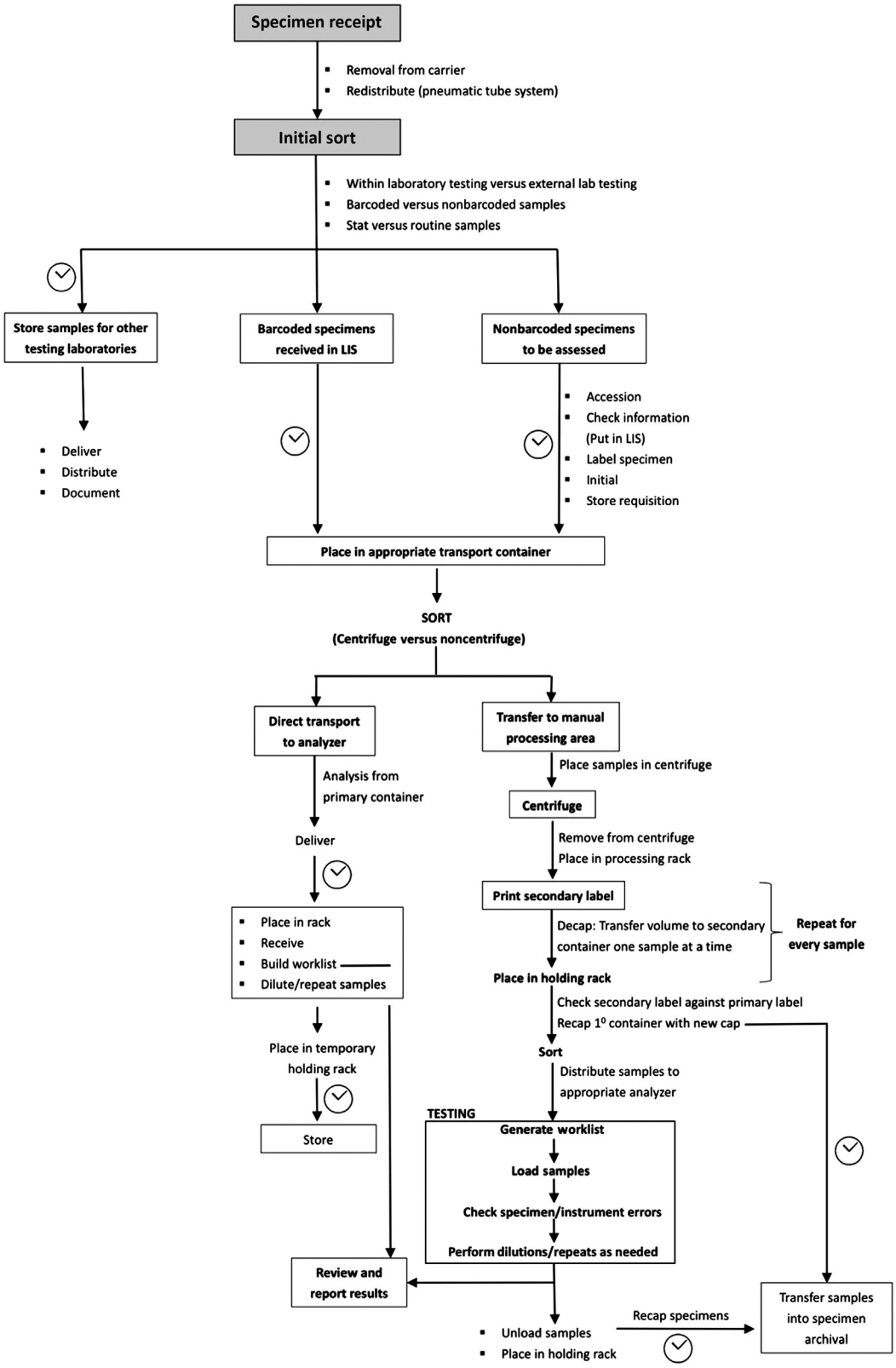


FIGURE 14.2 Simplified representative workflow integrating preanalytical, analytical, and postanalytical activities. Schematic clocks denote where workflow bottlenecks may occur.

vendors offering such automated solutions include bioMérieux Inc., Becton-Dickinson Diagnostics, and Copan Diagnostics. In addition, over the last decade, TLA systems targeting the clinical microbiology environment have been brought to market. Beyond exploiting automated barcode reading and a track-driven platform, incubators and digital equipment can be interfaced with the aforementioned microbiology and bacteriology specimen processing modules. BioMérieux Inc., BD Kiestra, and Copan Diagnostics all offer centralized automated clinical microbiology platforms. Further, a recent Q&A published in *Clinical Chemistry* reinforced the growing need and rationale for automation in clinical microbiology settings [10].

Fully integrated laboratory automation has also been met with success in clinical molecular laboratories, complementing work conducted in clinical microbiology. Various automation platforms offer barcode recognition of specimens, as well as automated nucleic acid extraction and polymerase chain reaction (PCR) amplification of DNA regions of interest. Laboratory automation focused on preanalytical DNA extraction, or PCR amplification and detection techniques are available through a number of vendors, including Abbott Laboratories, BD Diagnostics, bioMérieux Inc., Qiagen, Roche Diagnostics, and Thermo Fisher Scientific. However, much like automated chemistry and immunochemistry analyzers and TLA systems, specimen handling capabilities, tube-type flexibility, barcode detection, and laboratory automation software components are vendor-specific and should be considered when selecting an integrated platform.

Although many laboratories continue to look toward integrated automation solutions to reduce errors, improve efficiencies, and combat staff shortages, there are still areas that require automated preanalytical and centralized solutions. While mass spectrometry has successfully transitioned from a basic and translational tool to a clinical platform used for identification and/or quantification of small molecules and, more recently, proteins, specimen handling and management are still largely manual. Given the success of immunosuppressant and pain management testing by liquid chromatographic-mass spectrometric (LC-MS) techniques, centralized automated solutions may further integrate these methodologies into the clinical testing environment [11]. While vendors are pursuing automation-driven solutions with regard to mass spectrometry, including the automated review of mass spectra and chromatograms and interfacing of these data with the LIS, fully integrated automation for LC-MS platforms is still in its infancy.

Measurements of success

There are a number of ways through which the success and sustainability of subtotal or centralized laboratory automation can be evaluated postimplementation. Objective measures include comparison of specimen receipt with result TAT pre- and postimplementation, as well as the impact of automation on laboratory-derived errors. A number of reports have demonstrated improvements in TAT and reduction in laboratory errors postautomation. For many laboratories, TAT is defined as the time interval in which a specimen is received in the laboratory to the time a result is released or available to the provider. However, other TAT metrics may include the interval between order and result reporting, or collection and result reporting [7]. For the purpose of assessing the impact of automation of laboratory activities on overall performance, the interval from receipt in the laboratory to the releasing of results will be used and will be referred to as laboratory TAT. Laboratory TAT can be further stratified and defined based on test, priority, or patient population.

The positive impact of laboratory automation has been best demonstrated through the relationship between TAT reduction and improved patient outcomes in critical and emergent settings. Laboratory TAT is commonly used as a key performance indicator of laboratory performance, and delays in specimen resulting and unsatisfactory TAT events may not only result in complaints from providers, but also identify key areas for improvement. While recommendations and expectations regarding TAT may vary across institutions, a review by Hawkins has suggested a 90% successful laboratory TAT (specimen registration or receipt to result reporting) of $<$60 minutes for common laboratory tests as an initial TAT goal. In addition, physician and laboratory societies have provided recommendations regarding target TATs for stat samples collected in emergent settings. For example, both the American College of Cardiology/American Heart Association and the National Academy of Clinical Biochemistry (currently known as the American Association for Clinical Chemistry Academy) have recommended a laboratory TAT of $<$60 minutes for cardiac markers such as cardiac troponin (cTn) [12,13].

The thoughtful assessment of laboratory workflow and introduction of automation can assist in reaching quality metrics such as TAT. A case report from a large international hospital demonstrated an improvement in laboratory TAT postimplementation of a centralized laboratory automation system. A comparison of laboratory TAT both pre- and postautomation implementations was performed and a mean improvement of 27.7 $\pm$ 2.5 minutes in laboratory TAT for cTn samples, with TAT reductions also observed for nonemergent specimens [14]. Additional case reports have demonstrated improvements in laboratory TAT ranging from 30% to 50% following the incorporation of centralized automation systems [9]. However, as previously described, the benefits of automation in

improving TAT benchmarks are dependent on a clear understanding of the workflow and need of the laboratory.

In addition to reductions in the aforementioned preanalytical errors, labor and production costs and overall workload are additional metrics of automation success. Case reports from a number of institutions have demonstrated 50%−75% increases in sample volumes and workloads postimplementation of automation. These benchmarks were achieved without additional full-time employee (FTE) support. Laboratories have also observed 3−7 year payback on investment after implementation of integrated automation systems [9].

Oversight

With the growing incorporation and integration of complex and encompassing automation systems into the clinical laboratory environment, there is a growing need for standards and recommendations to ensure proper operation of the centralized platform. These concerns were raised in the 1990s by the Clinical Testing Automation Standards Steering Committee, who engaged the National Committee on Clinical Laboratory Standards [NCCLS; now known as the Clinical and Laboratory Standards Institute (CLSI)]. The NCCLS established a committee with an emphasis on automation to develop standards for laboratory automation. As a result, the CLSI has released a number of standard documents focused on both laboratory automation as well as LISs. Currently, approved standards are available for specimen containers (CLSI document AUTO01-A), barcode identification (AUTO02-A2), and communication with automated systems, instruments, devices, and information systems (AUTO03-A2), as well as nine other approved documents regarding laboratory automation. These documents can provide clarity to laboratories and vendors with regard to essential elements and components to be evaluated in the transition from manual to automated processes.

Limitations of automation

While centralized automation solutions can facilitate the standardization and streamlining of laboratory workflows, there are limitations in terms of the extent to which all steps in a process can become automated. Most centralized systems are currently unable to accommodate pediatric tubes (microtainers), resulting in the need for alternative and manual workflows to handle these specimens. Certain sample containers, including urine cups, may require upfront aliquoting before further handling on an automated line. Many laboratories also implement alternative workflows for the handling of rare and precious samples, such as body fluids. These may be processed offline prior to the analytical phase of testing. Thus all of these specimen types may require parallel infrastructure and processes to facilitate preparation and testing postspecimen receipt.

For centralized automation platforms, configurations typically involve a single preanalytical line that is connected to multiple analytical instruments. While most laboratories build in redundancy with regard to analyzers, the same is not the case for the preanalytical portion of a TLA system. Thus if there are track malfunctions or issues within the preanalytical modules of a centralized system, manual workflows would be required to continue routine operations. Further, centralized systems require at least some degree of downtime to replace reagents, reconcile software issues, or dispose of waste, among others. Depending on the degree of downtime for a preanalytical system, manual processes may be required to facilitate the continue management of emergent or stat samples. It is noted that, similar to analytical instruments, all of the components within a centralized system need to be properly maintained for optimal efficiency. This is further elaborated on in the following section.

Ongoing monitoring

An important consideration in the incorporation of centralized laboratory automation systems is the ongoing evaluation of its integrated functionality. While aforementioned metrics such as TAT and instrument downtime are important indicators of performance, other measures may also identify latent issues and improve overall workflow efficiency. Preanalytical processes can be evaluated for aliquot errors as well as transport efficiency of specimens throughout the preanalytical areas of the laboratory. Further, technical laboratory staff are well versed in the evaluation and assessment of analytical instrumentation, both via reviewing quality control material and performing proficiency testing challenges. In addition, automated analyzers produce alarms to indicate instrument-related issues; these may include temperature sensors, identification of a clog in an analyzer probe, or flagging an expired reagent. Postanalytical processes may be evaluated in several ways, including the review of autoverification rules and results reported to the hospital information system. As laboratories continue to introduce centralized automation into the workplace, understanding the pre- and postanalytical components is equally as important.

Maintenance performance on centrifuge and aliquot modules, monitoring the causes of defaulted samples that require manual follow-up, and identification of trends in preanalytical instrumentation performance can all serve as ways of proactively assessing centralized systems. As the preanalytical, analytical, and postanalytical phases of testing are consolidated in a centralized laboratory

automation system, expertise in the operation and ongoing evaluation of these components is critical to ensuring the benefits of integrated laboratory automation.

From a staff perspective, a survey of laboratory technologists, clinical laboratory scientists, and senior medical technologists highlighted frustrations centered around inadequate training for integration of laboratory automation system operation and troubleshooting [15]. Feedback from staff illustrated the need to ensure sufficient initial and ongoing training for all components of the centralized automation platform. Other concerns included issues with instrument downtime and perceptions regarding capacity, further supporting the need to utilize workflow analysis and establish realistic expectations for automation performance.

Future perspective

As TLA systems become more commonplace in clinical laboratory environments, the resulting question is then, "Where do we go from here?" Technology will continue to advance, offering more functionality and capacity to accommodate the growing responsibilities and diminishing resources given to the clinical laboratory. In a 2014 *Clinical Laboratory News* article, Dr. Robin Felder has postulated the future growth of automation, predicting further advances and expansions into the areas of specimen transportation, biobanking, cell-based assays, and home-based monitoring [5]. As processes continue to be streamlined and standardized via automation, the clinical laboratory will continue to play a central role in healthcare.

Conclusion

Over the last several decades, laboratory automation has progressed from CFA to centralized and total automation platforms, integrating many functions executed by the clinical laboratory. Automation platforms have become more prominent in laboratory settings and have become well accepted approaches of improving laboratory service, ensuring patient safety and accommodating cost reductions and staff shortages. As laboratories continue to "do more with less," automation will continue to be implemented and introduced into the laboratory environment. However, as demonstrated, the success of automation is dependent on understanding both existing workflows and expectant needs.

References

[1] CAP Today, College of American Pathologists. <https://www.captodayonline.com/2017/ProductGuides/09-17_CAPTODAY_LabAutomation.pdf> (accessed 23.01.19).

[2] S.E. Melanson, N.I. Lindeman, P. Jarolim, Selecting automation for the clinical chemistry laboratory, Arch. Pathol. Lab. Med. 131 (2007) 1063–1069.

[3] Clinical and Laboratory Standards Institute, Laboratory Automation: Bar Codes for Specimen Container Identification: Approved Standard. AUTO02-A2, CLSI, Wayne, PA, 2006.

[4] J.W. Holman, T.E. Mifflin, R.A. Felder, L.M. Demers, Evaluation of an automated preanalytical robotic workstation at two academic health centers, Clin. Chem. 48 (3) (2002) 540–548.

[5] R.A. Felder, Advances in clinical laboratory automation, Clin. Lab. News. (2014).

[6] J.C. Boyd, C.D. Hawker, Automation, in: C.A. Burtis, D.E. Bruns (Eds.), Tietz Fundamentals of Clinical Chemistry and Molecular Diagnostics, seventh ed., Elsevier, 2015, pp. 254–272.

[7] C.D. Hawker, Laboratory automation: total and subtotal, Clin. Lab. Med. 27 (2007) 749–770.

[8] C.D. Hawker, S.B. Garr, L.T. Hamilton, J.R. Penrose, E.R. Ashwood, R.L. Weiss, Automated transport and sorting system in a large reference laboratory: part 1. Evaluation of needs and alternatives and development of a plan, Clin. Chem. 48 (10) (2002) 1751–1760.

[9] C.D. Hawker, W.L. Roberts, S.B. Garr, L.T. Hamilton, J.R. Penrose, E.R. Ashwood, et al., Automated transport and sorting system in a large reference laboratory: part 2. Implementation of the system and performance measures over three years, Clin. Chem. 48 (10) (2002) 1761–1767.

[10] C.A. Burnham, W.M. Dunne Jr, G. Greub, S.M. Novak, R. Patel, Automation in the clinical microbiology laboratory, Clin. Chem. 59 (12) (2013) 1696–1702.

[11] D.A. Armbruster, D.R. Overcash, J. Reyes, Clinical chemistry laboratory automation in the 21st century-amat Victoria curam (victory loved careful preparation), Clin. Biochem. Rev. 35 (3) (2014) 143–153.

[12] J.L. Anderson, C.D. Adams, E.M. Antman, C.R. Bridges, R.M. Califf, J.D.E. Casey, et al., ACC/AHA 2007 guidelines for the management of patients with unstable angina/non–STelevation myocardial infarction—executive summary. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines, J. Am. Coll. Cardiol. 50 (7) (2007) 652–726.

[13] F. Apple, R. Jesse, L. Newby, A. Wu, R. Christenson, C. Cannon, et al., National Academy of Clinical Biochemistry. National Academy of Clinical Biochemistry and IFCC Committee for Standardization of Markers of Cardiac Damage Laboratory Medicine Practice Guidelines: analytical issues for biochemical markers of acute coronary syndromes, Clin. Chem. 53 (4) (2007) 547–551.

[14] C. Ialongo, O. Porzio, I. Giambini, S. Bernardini, Total automation for the core laboratory: improving the turnaround time helps to reduce the volume of ordered STAT tests, J. Lab. Autom. 21 (3) (2016) 451–458.

[15] C.W. Lam, E. Jacob, Implementing a laboratory automation system: experience of a large clinical laboratory, J. Lab. Autom. 17 (1) (2012) 16–23.

Self-assessment questions

1. The AutoAnalyzer I utilized the following analytical principle:
   a. centrifugal analysis
   b. TLA
   c. whole blood analysis
   d. CFA
2. The major distinction between subtotal and total laboratory automation is _____.
   a. subtotal automation platforms lack centrifugation capabilities
   b. subtotal automation platforms only identify barcoded specimens
   c. subtotal automation platforms do not have a connected track system
   d. subtotal automation platforms refer exclusively to the analytical phase of the total testing process
3. The typical barcode format used in clinical laboratory environments is _____.
   a. Code 39
   b. Code 128
   c. Code 28
   d. Codabar
4. Laboratory automation may reduce laboratory-associated errors by as much as _____.
   a. 10%
   b. 20%
   c. 50%
   d. 80%
5. What criteria are considered when evaluating TAT of automated instrumentation?
   a. The total number of samples tested in the laboratory
   b. The number of test requests associated with each specimen
   c. Instrument type and complexity
   d. All of the above
6. Which analytical testing area does not have integrated automation capabilities?
   a. Mass spectrometry
   b. Urinalysis
   c. PCR analysis
   d. Bacteriology
7. Which of the following is not a metric of automation success?
   a. Improved laboratory TAT
   b. Increased requirement for FTEs
   c. Increased laboratory volumes
   d. None of the above

Answers

1. d
2. c
3. b
4. d
5. d
6. a
7. b

Chapter 15

Laboratory regulations and compliance

Stanley T. Leung[1] and Jason Y. Park[2]

[1]*Incyte Diagnostics, Seattle, WA, United States,* [2]*Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX, United States*

Learning objectives

After reviewing this chapter, the reader should be able to:

- Discuss certification and accreditation of clinical laboratories.
- Discuss the importance of CLIA in the operation of clinical laboratories.
- Describe HIPAA and how its standards are applied in management of clinical information.
- Discuss adherence to OSHA regulations in the clinical laboratory.
- Describe regulations of laboratory billing and reimbursements.

Introduction

Operating a clinical laboratory requires both scientific and management expertise. In the United States, specific laws at the federal and state levels guide the analytical testing as well as the business operations of laboratories. This chapter will outline and discuss the most important regulations for clinical laboratories. Regulations affecting clinical laboratories will be broadly reviewed in the categories of quality, privacy, worker safety, and business practices.

Regulatory responsibility in the laboratory

Under federal law, laboratory directors have the overall responsibility of administration, operation, and employee competence. Within some healthcare organizations, some oversight may be assigned to a compliance officer, who monitors adherence to applicable regulations. In 1997 the US Department of Health and Human Services (HHS) Office of the Inspector General (OIG) issued a model compliance plan that was intended to address how laboratories should conduct their business regarding marketing and billing practices [1]. This model compliance plan contained the description of responsibilities for an individual in monitoring marketing and billing practices involving payment arising from federal programs. In practice, these responsibilities have generally expanded to include all billing practices, as well as compliance with other regulations. Laboratory directors have responsibilities for not only ensuring the quality of testing but also complying with appropriate business practices. The importance of the laboratory director is emphasized when noting that directorship involves two of the top 10 serious *condition-level* deficiencies cited by the Centers for Medicare and Medicaid Services (CMS) in 2018 (Table 15.1) [2].

Federal regulations

Clinical Laboratory Improvement Amendments of 1988

The federal regulation most important to clinical laboratory operations is the Clinical Laboratory Improvement Amendments of 1988 (CLIA '88 or CLIA). The original statute adopted by Congress can be found incorporated into the United States Code (USC) at 42 USC 263a [3]. Congress directed the development of "standards to assure consistent performance by laboratories [who have been] issued a certificate." It is these "standards" that have been developed into the regulations issued by CMS.

CLIA is described as "amendments" since the original Clinical Laboratory Improvement Act had been passed in 1967. The original statute only applied to laboratories performing tests across state borders (interstate commerce) and affected mostly hospital-based and independent laboratories. The 1967 law specifically exempted physician office laboratories (POL). However, during the 1980s, there were a series of media stories identifying errors in laboratory testing that led to misdiagnosis, delayed diagnosis, and/or patient injury. Although the stories in the media focused on cervical cytology screening, drug

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00015-6

TABLE 15.1 Top 10 condition-level[a] deficiencies from Centers for Medicare & Medicaid Services surveys (2018).

Regulatory cite	Deficiency	Labs deficient	Percent deficient (%)
493.1403	Laboratory must have a director who meets the qualification requirements of 403.1405 and provides overall management and direction in accordance with 493.1407	417	2.48
493.1441	Laboratory must have a director who meets the qualification requirements of 493.1443 and provides overall management and direction in accordance with 493.1445	266	1.58
493.1250	Laboratory that performs nonwaived testing must meet the applicable analytic systems requirements in 493.1251 through 493.1283	236	1.41
493.1421	Laboratory must have sufficient number of individuals meeting the qualification of 493.1423 to perform the functions specified in 493.1425 for the volume and complexity of tests performed	185	1.10
493.1409	Laboratory must have a technical consultant who meets the qualification requirements of 493.1411 and provides technical oversight in accordance with 493.1413	189	1.13
493.801	Each laboratory must enroll in a PT program that meets the criteria in subpart I and is approved by HHS. The laboratory must enroll for each specialty and subspecialty, and must test the samples in the same manner as patients' specimens	179	1.07
493.803	Each laboratory performing nonwaived testing must successfully participate in a PT program approved by CMS as described in subpart I of this part for each specialty, subspecialty, and analyte or test in which the laboratory is certified under CLIA	117	0.79
493.1487	Laboratory must have sufficient number of individuals who meet the qualification requirements concerning State licensure, if applicable, and the educational requirements for high-complexity personnel as defined in 493.1489	92	0.55
493.1215	If the laboratory provides services in the specialty of hematology, the laboratory must meet the requirements specified in 493.1230 through 493.1256; and 493.1269 and 493.1281 through 493.1299.	65	0.39
493.1447	Have a technical supervisor who meets the qualification requirements as defined in 493.1449 and provides technical supervision in accordance with 493.1451	59	0.35

CLIA, Clinical Laboratory Improvement Amendments; *CMS*, Centers for Medicare & Medicaid Services; *HHS*, Health and Human Services; *PT*, proficiency testing.
[a]*CMS laboratory deficiencies are categorized as standard-level and condition-level; condition-level deficiencies are a more serious category and require correction before CMS will issue a certificate.*

testing, and AIDS testing, Congressional debate began to evolve around broader quality issues requiring oversight throughout all aspects of the clinical laboratory. Importantly, CLIA '88 was enacted to apply to all clinical laboratories including POL and small private laboratories. In addition, CLIA '88 not only applies to clinical laboratories but also applies to any person (e.g., physician and medical assistant), performing testing on human specimens for the diagnosis, prevention, or treatment of health conditions. Although CLIA '88 is administered under CMS, it applies to all testing situations in the United States regardless of whether the patient is a Medicare beneficiary. CLIA '88 set minimal national standards for laboratory quality assurance, quality control (QC), and proficiency testing (PT). The final version of CLIA '88 was published and became effective in 1992.

Administrative provisions

The foundation of CLIA is that all tests performed on clinical specimens are subject to CLIA's provisions. CLIA also applies to testing sites ranging from limited POLs with a few types of tests to free standing commercial reference laboratories with thousands of types of tests. However, when the laboratory testing is not intended for diagnostic or therapeutic purposes, then the application of CLIA is less clear. Testing for research purposes is excluded, as is drug testing performed by employers and forensic testing. However, CLIA applies to all tests that are communicated to clinical care providers or entered in the patient's medical record. Under CLIA, analytical results from research laboratories cannot be communicated to patients or be used by physicians or other clinical care providers.

Administrative provisions of CLIA created a nongovernmental advisory committee, the Clinical Laboratory Improvement Advisory Committee (CLIAC), consisting of 20 members selected by the HHS plus members from the Centers for Disease Control and Prevention, CMS, and the Food and Drug Administration (FDA). CLIAC meets several times a year to consider issues referred to it by CMS and to develop recommendations in response to questions.

Certification process

The CLIA regulation recognizes that different tests require different levels of scrutiny depending on the level of complexity of the testing performed. The categories of tests are (1) waived tests; (2) tests performed by the physician or advanced practitioners using microscopy [provider-performed microscopy (PPM)]; (3) moderately complex tests; and (4) highly complex tests. This classification is critically significant, because it determines the level of oversight and regulatory compliance requirements including PT, QC, and personnel qualification. The FDA determines the test complexity category.

Waived tests

The original language of CLIA described waived tests as those "methodologies that are so simple and accurate as to render the likelihood of erroneous results negligible, or ... [those posing] no reasonable risk of harm to the patient if performed incorrectly." The waived tests have expanded over time and now include some molecular diagnostic tests (Table 15.2). In 2005 the FDA issued a draft recommendation to manufacturers who seek to have their in vitro devices (IVDs) classified as waived [4]. To be considered for waived classification, a testing device cannot require any preanalytical steps (i.e., no specimen processing). In addition, the manufacturer must analyze the effects of deliberately making errors in test performance. The test accuracy, including a description of a comparative method, must be provided, and no postanalytical calculations can be required. As a result, a person with no technical training should be able to produce the same results as those produced by a trained medical technologist.

A laboratory holding a Certificate of Waiver (meaning it is approved to perform waived tests only) needs to follow two basic rules: It must perform only the waived tests named in its application, and it must perform those tests

TABLE 15.2 Examples of Clinical Laboratory Improvement Amendments waived tests.

Test	Use	CPT code
Dipstick or tablet reagent urinalysis	Screening urine to monitor/diagnose various diseases	81002
Urine pregnancy tests by visual color comparison	Diagnosis of pregnancy	81025
Fecal occult blood	Detection of blood in feces from various causes	82270/ 82272
Blood glucose by glucose monitoring devices cleared by the FDA for home use	Monitoring of blood glucose levels	82962
Hemoglobin by copper sulfate—nonautomated	Monitors hemoglobin level in blood	83026
Ovulation tests by visual color comparison for human luteinizing hormone	Detection of ovulation	84830
Blood count; spun microhematocrit	Screen for anemia	85013
Erythrocyte sedimentation rate—nonautomated	Screening test for inflammatory activity	85651
Abbott i-STAT Chem8 + Cartridges	Measures multiple peripheral blood chemistry analytes	Multiple
Abaxis Piccolo Blood Chemistry Analyzer	Measures multiple peripheral blood chemistry analytes	Multiple
Alere I Influenza A&B Test	Detection of Influenza A and B viral nucleic acids using isothermal nucleic acid amplification technology	87502QW
Cepheid Gene Xpert Xpress System	Multiple RT-PCR assay for detection of Influenza A and B	87502QW
Streptococcus group A antigen by immunoassay with direct optical observation	Rapid detection of the group A streptococcus antigen for the diagnosis of infection	87880QW

Notes: An updated list of all waived tests is available on the CMS website: https://www.cms.gov/Regulations-and-Guidance/Legislation/CLIA/Downloads/waivetbl.pdf. *CPT*, Current procedural terminology; *FDA*, Food and Drug Administration.

in compliance with the manufacturer's instructions. Given that more than half of US laboratories are in the category of waived-testing laboratories, there is a significant incentive for a manufacturer to have its test classified as waived.

Provider-performed microscopy

These nine-specific manual microscopic examinations are by definition not "kits" or specific methods (Table 15.3). These tests are used to identify certain microorganisms or features, using bright-field or phase-contrast microscopy. They can only be performed by physicians or midlevel practitioners (e.g., advanced practice nurses, physician assistants, and nurse midwives). No requirements in the form of traditional QC exist, and no PT is required. However, PPM requires written policies and procedures for the assessment of personnel competency. Training and demonstrated competency must be documented. The regulation of PPM is a looser type of regulation based on the premise that physicians and midlevel practitioners interact directly with the patient, and therefore can correlate their microscopic findings with their clinical assessment. When PPM tests are performed in a hospital setting, laboratory staff may assist with or coordinate compliance, but generally, entities holding a PPM Certificate have little to do with the traditional laboratory. Although traditional QC and PT is not required, CLIA considers PPM a special type of moderate-complexity testing, and as a result, all waived testing can also be performed under a PPM certificate.

Moderate- and high-complexity testing

The distinction between moderate- and high-complexity tests is based on the characteristics of the testing process (Table 15.4). Most automated testing is classified as moderate complexity, whereas tests that are largely manual (electrophoresis, chromatography, microbiological procedures, and molecular pathology) are all high complexity. Initially, the classification as moderate- or high-complexity tests determined the QC procedures that the regulations required and the extent of test validation expected. These differences were dropped in a revision of the CLIA regulation issued in 2004, and the principal difference between moderate- and high-complexity testings is now in personnel qualifications of laboratory staff (Table 15.5). In addition to laboratory certification, there is individual certification with specific board organizations that are recognized by HHS for certifying individuals with a PhD in chemical, physical, biological, or clinical laboratory science (Table 15.6). Understanding the personnel qualifications of laboratory staff is critical to avoiding regulatory action. As noted in the *condition-level* deficiencies cited by CMS, personnel issues are six of the top 10 most common deficiencies (Table 15.1).

To be certified to perform moderate- or high-complexity testings, a laboratory must possess a Certificate of Accreditation. Unlike the Certificate of Waiver, which imposes very limited requirements on the waived laboratory, the Certificate of Accreditation imposes significant requirements. Certification requires a biennial inspection by an organization authorized by CMS to inspect and approve laboratories as meeting standards complying with CLIA requirements. These standards include successful participation in PT; compliance with QC, maintenance, and record keeping requirements; demonstration that the laboratory staff meets CLIA educational and certification requirements; and payment of a fee for issuance of the certificate (Table 15.7).

TABLE 15.3 Provider-performed microscopy tests.

Test	CPT code
Wet mounts, including vaginal, cervical, or skin specimens	Q0111
Potassium hydroxide preparations	Q0112
Pinworm examinations	Q0113
Fern test	Q0114
Postcoital direct, vaginal, or cervical mucous examination	Q0115
Microscopic urinalysis	81015
Fecal leukocyte examination	89055
Semen analysis	G0027
Nasal smears for eosinophils	89190

TABLE 15.4 Food and Drug Administration criteria for categorizing test complexity.

Category	Lowest complexity	Highest complexity
Technical and scientific knowledge	Minimal knowledge	Specialized knowledge is essential
Training and experience needed for all phases of testing process	Minimal training	Specialized training or substantial experience is essential
Reagent and material preparation	Reagents and materials are stable, reliable, and/or prepackaged	Reagents and materials require special handling or require manual steps such as gravimetric or volumetric measurements
Characteristics of operational steps	Automatically executed or easily controlled	Close monitoring or control are required (e.g., precise temperature control, accurate pipetting, and extensive calculations)
Calibration, quality control, and proficiency testing materials	Stable and readily available calibrators	Require special handling
Test system troubleshooting and equipment maintenance	Automatic or self-correcting; requires minimal judgement	Manual; requires decision-making to resolve most problems
Interpretation and judgement	Minimal interpretation or judgement	Extensive interpretation and/or judgement required

TABLE 15.5 Staffing requirements for laboratories performing moderate- and high-complexity testings.

Position	Moderate-complexity testing	High-complexity testing
Director	MD/DO, board certified in pathology	MD/DO, board certified in pathology
	MD/DO/DPM, and one of • 1-year experience directing/supervising • 20-h lab CME • 20-h lab training in residency	MD/DO/DPM, and one of • 2-year experience directing/supervising • 1-year lab training in residency
	PhD,[a] certified by HHS recognized board	PhD, certified by HHS recognized board
	PhD and 1-year experience	PhD and all of • previously served as high-complexity director on or before December 31, 2002 • 2-year training/experience • 2-year experience directing/supervising
	MS[b] and all of • 1-year training/experience • 1-year supervisory experience	
	BS[c] and all of • 2-year training/experience • 2-year supervisory experience	
	Previously qualified before February 28, 1992	Previously qualified and serving on or February 28, 1992
Clinical Consultant	MD/DO, board certified in pathology	MD/DO, board certified in pathology
	MD/DO/DPM and one of • 1-year experience • 20-h lab CME • 20-h lab training in residency	MD/DO/DPM and one of • 2-year experience directing/supervising • 20-h lab CME • 1-year lab training in residency
	PhD, certified by HHS recognized board	PhD, certified by HHS recognized board

(*Continued*)

TABLE 15.5 (Continued)

Position	Moderate-complexity testing	High-complexity testing
Technical consultant	MD/DO, board certified in pathology	MD/DO, board certified in pathology
	MD/DO/DPM, 1-year training/experience in specialty	MD/DO/DPM, 1-year training/experience in specialty
	PhD, 1-year training/experience in specialty	PhD, 1-year training/experience in specialty
	MS, 1-year training/experience in specialty	MS, 2-year training/experience in specialty
	BS, 2-year training/experience in specialty	BS, 4-year training/experience in specialty
General supervisor	Not required	Qualify as director/technical consultant for high-complexity testing
		MD/DO/DPM/PhD/MS/BS, 1 y experience
		AA or equivalent and 2-year experience
		Previously qualified as general supervisor on or before February 28, 1992
		Served as a general supervisor on or before September 1, 1992 with 2-year training/experience and one of • graduated medical laboratory or clinical lab program • HS diploma and military training to serve as medical lab specialist
		Served as a general supervisor on or before September 1, 1992 and all of • HS diploma or equivalent • 10 y training/experience between 1982 and 1992
Testing personnel	MD/DO/DPM/PhD/MS/BS/AA	MD/DO/DPM/PhD/MS/BS/AA
	HS diploma and military training to serve as medical lab specialist	On or before April 24, 1995, HS diploma and one of • graduate of approved lab training program • military training to serve as medical lab specialist
	HS diploma or equivalent and appropriate training or experience	

AA, Associate in arts; BS, Bachelor of Science; *CME*, Continuing Medical Education; *DO*, Doctor of Osteopathy; *DPM*, Doctor of Podiatric Medicine; *HHS*, Health and Human Services, *HS*, High School; *MD*, Medical Doctor; *MS*, Master of Science; *PhD*, Doctor of Philosophy.
[a]*PhD in chemical, physical, biological, or clinical laboratory science.*
[b]*MS in clinical laboratory science, medical technology, or chemical, physical, or biological science.*
[c]*BS in medical technology or chemical, physical, or biological science.*

TABLE 15.6 US Department of Health and Human Services approved certification boards for PhD laboratory directors of high-complexity testing.

Abbreviation	Certification board
ABB	American Board of Bioanalysis
ABCC	American Board of Clinical Chemistry
ABFT	American Board of Forensic Toxicology
ABHI	American Board of Histocompatibility and Immunogenetics
ABMGG	American Board of Medical Genetics and Genomics
ABMLI	American Board of Medical Laboratory Immunology
ABMM	American Board of Medical Microbiology
NRCC	National Registry of Certified Chemists

Notes: A medical doctor, doctor of osteopathy, doctor of podiatric medicine, or PhD with the appropriate experience can be a high-complexity laboratory director. A PhD laboratory director must hold an earned doctoral degree in a chemical, physical, biological, or clinical laboratory science from an accredited institution and be certified and continue to be certified by a board approved by HHS. ABFT and NRCC also certify nondoctoral individuals.

TABLE 15.7 Clinical Laboratory Improvement Amendments certificate fees, as of 2017.

Type of lab	Number of specialties	Annual test volume	Biennial certificate fee
Waived	N/A	N/A	$180
PPM	N/A	N/A	$240
Low volume A	N/A	Less than 2000	$180
Schedule A	Three or fewer	2000–10,000	$180
Schedule B	Four or more	2000–10,000	$180
Schedule C	Three or fewer	10,001–25,000	$516
Schedule D	Four or more	10,001–25,000	$528
Schedule E	N/A	25,001–50,000	$780
Schedule F	N/A	50,001–75,000	$1,320
Schedule G	N/A	75,001–100,000	$1,860
Schedule H	N/A	100,001–500,000	$2,448
Schedule I	N/A	500,001–1,000,000	$7,464
Schedule J	N/A	>1,000,000	$9,528

PPM, Provider-performed microscopy.

TABLE 15.8 Accreditation organization with deeming authority under Clinical Laboratory Improvement Amendments.

AABB American Association for Laboratory Accreditation, American Osteopathic Association/Health Facilities Accreditation Program, American Society for Histocompatibility and Immunogenetics
COLA College of American Pathologists Joint Commission

Notes: Only American Association for Laboratory Accreditation, American Osteopathic Association/Health Facilities Accreditation Program, College of American Pathologists, and Joint Commission provide accreditation for all the major specialties of histocompatibility, microbiology, diagnostic immunology, chemistry, hematology, immunohematology, pathology, and cytogenetics. Only two states are exempt from Clinical Laboratory Improvement Amendments: Washington and New York. *AABB*, American Association of Blood Banks; *COLA*, Commission on Office Laboratory Accreditation.

Inspecting organizations

The laboratory inspection, which must be performed before initiation of laboratory operation and at 2-year intervals thereafter, may be performed by a state agency responsible for public health or by a private organization approved by CMS for conducting accreditation inspections. Approved entities are referred to as *deemed agencies*, because they have been deemed to represent CMS for the purpose of accreditation compliance. There are multiple organizations with deemed status (Table 15.8): American Association of Blood Banks (now called AABB), Commission on Office Laboratory Accreditation (now called COLA), American Association for Laboratory Accreditation (A2LA), American Osteopathic Association, American Society for Histocompatibility and Immunogenetics, College of American Pathologists (CAP), and the Joint Commission for Accreditation of Healthcare Organizations (Joint Commission). Each deemed organization has specific areas of the laboratory, which they can inspect. Only A2LA, American Osteopathic Association, CAP, and the Joint Commission cover all of the major laboratory areas recognized by CLIA.

Laboratory inspections have a fee (payable to the inspecting organization), which is separate from the certificate fee (payable to CMS). There are subtle differences in the perspective of the deemed organizations' inspections compared with inspections conducted by governmental agencies. The CLIA certificate fee depends on the

number of different specialty types of tests (chemistry, hematology, coagulation, microbiology, transfusion services, etc.) and the test volume for a given laboratory.

Proficiency testing

A CLIA certified laboratory must participate in PT using an approved external provider of proficiency test materials. PT, when it is conducted for the purposes of compliance with the provisions of CLIA, is evaluated according to criteria defined in the regulations. In general, the CLIA PT requirements are less stringent than PT evaluations conducted for quality assurance purposes. Not all tests have the same PT requirements; CLIA specifically lists the tests that require PT as well as the limits of acceptability. PT must be performed in a manner identical to the way patient samples are tested, and communication with other laboratories regarding proficiency test results is prohibited. PT must be performed by staff who routinely perform patient testing, and no extra calibration of instruments or duplicate testing of proficiency samples are allowed. Documentation should be maintained to support the proper manner of testing.

For tests requiring PT, there are two types of analytes: regulated and nonregulated. For regulated analytes, CLIA specifies that three sets of five proficiency challenges must be completed by accredited laboratories each year (Table 15.9). Although achieving acceptable results is not typically challenging, the consequences of repeated failures can be dire, and even more severe consequences may result from deliberate violation of the prohibition of "cheating" (e.g., having proficiency samples run at another laboratory or in collaboration with others regarding results of PT). When a set of five specimens is received, the laboratory must produce acceptable results on at least four of the five to receive a satisfactory grade for that challenge. In other words, an 80% level of acceptability is satisfactory. If an unsatisfactory result ($<$80% acceptable) is obtained, then the laboratory must achieve at least 80% acceptability on the next two subsequent challenges for the same analyte, or the laboratory's PT performance is considered unsuccessful. Unsuccessful PT results may eventually prompt an investigation of the laboratory, with the possibility of suspension of certification until correction and improvement is proven. During a suspension, a laboratory may be prohibited from operating or may be required to refrain from testing in the area that was unsuccessful. Suspensions are not automatic or instantaneous but depend on the outcome of an investigation.

Patient test management

CLIA contains many provisions regulating the pre- and postanalytical phases of testing. This section includes requirements such as documentation of physician orders for tests performed, the availability of an up-to-date procedure manual, records of instrument maintenance and repair, and reports of test results that include a reference range and any pertinent patient and/or specimen information, as well as standards for records retention.

Quality control

The first version of the CLIA regulations contained QC requirements that reflected the long-standing laboratory tradition of running two liquid controls with every run of patient samples. Over time, however, it became apparent that analytical inaccuracy was not the most common cause of laboratory error, and an updated version of QC regulations was published in 2005. One driving force behind this change was the increased use of single-use testing devices, in which the traditional statistical sampling approach for QC did not apply. In addition, test device manufacturers have developed electronic controls capable of monitoring various elements of the test instrumentation, providing assurance that test components are operating within defined specifications.

While retaining the two-controls-per-run practice as an acceptable QC practice, CMS has allowed laboratories to document the effectiveness of alternative approaches to QC such as electronic checks, use of patient means, and other nontraditional approaches to ensure the accuracy of testing. Overall, CMS is attempting to encourage the development and use of a systems approach to quality assurance in place of the traditional focus on the analytical phase of testing. An example of an innovative practice is *Individualized Quality Control Plan*, an alternative QC option under CLIA, which permits laboratories some customization in their QC practice.

Personnel standards

Compliance with CLIA further requires that the laboratory be directed by a qualified individual, with qualification being determined by academic degree and experience (Table 15.4). Some states have set additional requirements for laboratory staff qualifications, and state requirements must be met when they are more stringent than CLIA.

Enforcement

Ultimately, failure to comply with CLIA regulations is a criminal offense, carrying the possibility of conviction and punishment. In the laboratory profession, there have been cases in which laboratory owners and top managers have spent time in jail, but historically such cases have involved fraudulent billing, not QC or PT failures. The most serious penalty facing a laboratory charged with CLIA violations is a bar to participation in Medicare business.

TABLE 15.9 Clinical Laboratory Improvement Amendments-regulated analytes requiring proficiency testing[a]

Lab specialty	Tests	
Microbiology	Acid fast stain Aerobic/anaerobic culture and identification Antibiotic susceptibility testing Direct bacterial antigen detection Direct viral antigen detection Gram stain	Identification of parasites Mycobacteriology identification Mycobacteriology susceptibility Mycology culture and identification Presence or absence of parasites Viral isolation and identification
Diagnostic immunology	Alpha-1 antitrypsin Alpha fetoprotein Antinuclear antibody Antistreptolysin O Antihuman immunodeficiency virus Complement C3 Complement C4 Hepatitis B surface antigen	Hepatitis B core antibody Hepatitis B e-antigen Immunoglobulins, total (IgA, IgG, IgM, and IgE) Infectious mononucleosis Rheumatoid factor Rubella Syphilis serology
Chemistry	Alanine aminotransferase Albumin Alkaline phosphatase Amylase Aspartate aminotransferase Bilirubin, total Blood alcohol Blood gases: pH, pCO_2, and pO_2 Blood lead Calcium, total Carbamazepine Chloride Cholesterol, total Cholesterol, HDL Cortisol Creatine kinase, total Creatine kinase, isoenzyme Creatinine Digoxin Ethosuximide Free thyroxine Gentamicin Glucose Human chorionic gonadotropin	Iron, total LDH LDH isoenzymes (LDH1/LDH2) Lithium Magnesium Phenobarbital Phenytoin Potassium Primidone Procainamide Quinidine Sodium T3 uptake Theophylline Thyroid stimulating hormone Thyroxine, total (T4) Tobramycin Total protein Triglycerides Triiodothyronine (T3) Urea nitrogen Uric acid Valproic acid
Hematology	Cell identification Erythrocyte count Fibrinogen Hematocrit Hemoglobin	Leukocyte count Partial thromboplastin time Platelet count Prothrombin time WBC differential
Immunohematology	ABO Group D (Rho) typing Unexpected antibody detection	Compatibility testing Antibody identification

AST, Aspartate aminotransferase; *LDH*, lactate dehydrogenase.

[a]*This list of tests is current as of February 2019; there are proposals to add and delete tests from this list.*

Patient privacy

The *Health Insurance Portability and Accountability Act* (HIPAA) of 1996 was passed into law to ensure that an employee could not be denied health insurance (e.g., because of a preexisting condition) if he or she changed employers. At the same time, it created a single electronic standard; it was recognized that this electronic standard would increase the privacy risk of unauthorized persons gaining access to private health-related information.

Therefore strict privacy provisions were entered into the regulatory framework, which would cover all healthcare providers. Because laboratories are the largest single producer of data, the burden of protecting that data falls heavily on the staff who manage and operate laboratory information systems. HIPAA's protections fell into two categories: the privacy standard and the security standard.

Privacy standard

Healthcare providers are required to provide a statement explaining how they handle confidential patient information [protected health information (PHI)]. Eighteen specific identifiers including name, birth date, admission date, and discharge date are all described as PHI by HIPAA. This manner of handling the PHI must be in compliance with HIPAA, and the provider must provide this statement to each patient and obtain a signature acknowledging receipt. An organization with multiple providers such as hospitals can merge multiple HIPAA notices into a single statement with a single patient signature. Providers that do not have direct contact with patients (i.e., clinical laboratorians) are not required to obtain such a signature; however, they need to demonstrate a good faith effort to ensure that their HIPAA statement reaches the patients whom they serve. The HIPAA privacy standard further defines when PHI can be shared without the permission of the patient. The major categories of sharing PHI without permission are treatment, operations, and payment. The treatment category is what authorizes reporting of laboratory results to physicians, nurses, consultants, and any other provider treating the patient. Laboratories are expected to use reasonable judgment in balancing the benefit of communicating a result for the patient's treatment with the obligation to protect the patient's privacy. For example, if a telephone caller identifies himself or herself as a nurse or physician in need of laboratory data on a certain patient, the laboratory may provide that information without written documentation of the request or proof of the caller's identity. The operations category allows laboratory staff to review the test results or the clinical record of any patient for whom tests are done. Similarly, because QC and certification are operational requirements of a laboratory, access to patient results by an inspecting entity, such as CAP or Joint Commission, is allowed. Finally, the payment authorization allows passing PHI to an outside billing agency for the purpose of seeking payment for services rendered.

Any other release or PHI requires either the authorization of the patient or record keeping as an unauthorized release. Provision of records in response to a subpoena or in compliance with health department requirements (e.g., reporting of infectious disease cases) must be recorded, as well as any accidental or erroneous release of patient records. This might occur if a patient report is sent to the wrong physician, for example. The patient is entitled to request an accounting of all unauthorized releases of his or her data. The intent of this provision of HIPAA was so that patients could be informed of accidental or misdirected releases of their PHI. Compliance with the privacy standard is typically achieved with the following approaches:

- control of paper containing PHI (avoiding commingling with general trash);
- shredding of documents containing PHI;
- limiting the inclusion of PHI on reports and worksheets to the minimum necessary to achieve the purpose of such reports; and
- recording of all releases of information that is not specifically related to treatment, operations, and payment.

The security standard

The security aspects of HIPAA require creating and enforcing policies that provide physical security such as using passwords, maintaining computer firewalls, and investigating attempted unauthorized entry into an electronic database. This is generally the responsibility of information systems personnel and is very similar to the requirements commonly in force for the protection of finance- and accounting-related systems. Generally, once these provisions are put into place, they continue to operate to meet the security standard. Because the requirements of HIPAA cover all healthcare providers and not specifically laboratories, there is typically a unified effort to comply and to find strategies that work for all elements within a healthcare operation.

Health Information Technology for Economic and Clinical Health Act

Enacted in 2009, the Health Information Technology for Economic and Clinical Health (HITECH) Act provides additional civil and criminal enforcement of the HIPAA requirements for patient privacy and healthcare data security. HITECH specifies actions for reporting HIPAA violations; for example, when more than 500 patients are affected by a data breach (unauthorized disclosure of PHI), not only are patients notified, but HHS and the news media must also be notified. Under HITECH, the civil penalties can be as high as $1.5 million for repeat violations.

Genetic privacy

Genetic data are not one of the 18 specific PHI identifiers under HIPAA. Because genetic data are not considered PHI, genetic data are protected under Federal law only

when it is associated with one or more of the 18 specific PHI identifiers. Once genetic data are deidentified, it is no longer protected under Federal law. Although there is no current Federal requirement for protecting deidentified genetic data, clinical laboratories should consider the future risk to patient privacy if deidentified data are reidentified. Reidentification of genetic data has been demonstrated using publicly available resources such as social media, phone books, and genealogy records. An individual patient's genetic data provide insights into the genetic data of all biologic relatives. Thus risks to the genetic privacy of one individual are transferred to all biologic relatives.

Biobanking and providing specimens for research

Clinical laboratories may serve as resources for biologic materials in research. After patient samples are used for clinical testing, the remaining (residual) material is considered waste and may qualify for future research. Biobanks may also contain materials specifically collected for research. The main federal regulations to consider include institutional review board (IRB) oversight under the Department of HHS' Office for Human Research Protections, patient privacy, and determining whether the proposed use of the materials is considered human subjects research under the "Common Rule." The Federal "Common Rule" is the standard of ethics for human subjects research in the United States and defines the actions of IRB oversight. The establishment and ongoing activities of a residual human specimen biobank need to be reviewed and approved by a local IRB. In addition, any project that intends to use specimens from the biobank needs separate IRB oversight and approval.

Occupational Safety and Health Administration regulations

The federal Occupational Safety and Health Administration (OSHA) is responsible for workplace safety and health issues. OSHA focuses on workplaces in general and not on healthcare or clinical laboratories in particular. The underlying principle is protection of workers and not protection of patients or quality of the services provided. If a laboratory were operated without a single human worker, then OSHA would not apply to that laboratory. Two specific OSHA rules are pertinent to clinical laboratories: bloodborne pathogens and chemical hygiene regulations.

Bloodborne pathogens

There are federal laws that contain detailed protections that must be implemented to prevent transmission of infectious disease to workers through exposure to blood or blood-containing products. The regulations are referred to as "The Bloodborne Pathogen Regulation" and set standards in the following areas:

- personal protective equipment such as fluid-resistant lab coats, eye protection, and gloves;
- engineering controls such as needle-stick prevention devices, hand-washing sinks, and barriers used when opening tubes containing blood;
- housekeeping, including decontamination of workbenches and equipment, and disposal of infectious waste;
- employer-provided hepatitis B vaccination;
- communication and training; and
- record keeping of exposure incidents and review of causation and potential preventive measures.

The Bloodborne Pathogen Regulation underlies the protective measures that are part of every accreditation inspection. Although stool and urine can contain equally dangerous infectious organisms, they are not covered by this rule.

Chemical hygiene rule

The Chemical Hygiene Rule is a federal regulation intended to prevent exposure-related injury to employees. The focus is on toxic or harmful materials such as carcinogens; caustic, irritating, and combustible materials; explosive materials; and compressed gases. The regulation requires an organization to designate a chemical hygiene officer and, if appropriate, a chemical hygiene committee. This person or committee is responsible for the creation and maintenance of a chemical hygiene plan and development of a list of dangerous chemicals on hand, the amount used annually, and plans for protection of employees in the event of an accidental spill. However, when chemicals are commercially prepared and all of the reagents needed to conduct a test are contained in a kit, then OSHA considers them as having no potential for employee exposure, and they are thus outside the scope of the regulation.

Reimbursement regulations: fraud and abuse

When compared to compliance with CLIA and OSHA regulations, handling reimbursement under the Medicare and Medicaid programs is a more complex challenge. The issue of fraudulent billing arose in the mid-1990s in the context of rapidly escalating healthcare costs. Medicare is generally the government program, providing healthcare benefits for the elderly, and Medicaid is the government program providing healthcare benefits for the poor.

These were both created in 1966. Initially opposed by physicians, Medicare payments were divided into Part A (healthcare facility technical fees) and Part B (physician professional fees). Because reimbursement was initially based on what was actually billed, healthcare providers and facilities virtually controlled reimbursement rates. The programs were quickly recognized as an opportunity for significant income by all healthcare providers, including laboratories. As a result, government costs began to rise more rapidly than initially projected, so the Health Care Finance Administration (now called CMS) began implementing techniques to control Medicare spending.

For laboratory services, a major change occurred in 1984, when payment for laboratory testing performed on Medicare patients began reimbursement on the basis of a fee schedule amount instead of a "reasonable cost" basis. This initial fee schedule was based on prevailing charges submitted by laboratories during 1982−83 and then discounted to reflect the discounts commonly offered by laboratories. Despite the promise that this fee schedule would be increased for inflation, the increase was deferred or eliminated more than 50% of the time between 1985 and the present, while, at the same time, Congress mandated "National Limitation Amounts," which effectively decreased the payment to laboratories even further. More than 20 years after implementation of the fee schedule and despite many changes in testing technology, Medicare is still reimbursing for laboratory testing under its original fee schedule. As a result, reimbursement for a peripheral blood thyroid stimulating hormone (TSH) test is still approximately seven times the reimbursement for glucose. At the other end of the spectrum, newer immunoassay tests, if they do not have specific billing codes assigned, must be submitted as a nonspecific "Immunoassay, Other," which is reimbursed substantially below the cost of testing. As a result, CMS receives many complaints from the laboratory profession that Medicare underreimburses for certain tests. An investigation by the Institute of Medicine found that Medicare laboratory payment policy is badly flawed [5].

In 2014 Congress passed the Protecting Access to Medicare Act (PAMA), which significantly changes how CMS pays for clinical laboratory tests under the Clinical Laboratory Fee Schedule (CLFS). PAMA requires CMS to establish a new market-based payment system; these new payment rates apply to most clinical tests, including POLs. This market-based system requires certain laboratories, including physician office-based labs, independent labs, and hospital outreach labs, to collect and report private payor payment and test volume data to CMS. For each laboratory test that is reported for a data collection period, CMS calculates a weighted median payment for the test. As a result, in 2018, the payment amount for most tests is equal to the weighted median of the *surveyed* private payor rates. Payment rates under the private payor rate-based CLFS are updated every 3 years. If the Secretary of HHS determines that an applicable laboratory has failed to report or made a misrepresentation or omission in reporting information regarding a clinical diagnostic laboratory test under PAMA, a civil monetary penalty of up to $10,000 per day for each failure to report or each such misrepresentation or omission can be assessed.

The original intent of PAMA was to ensure that the new payment rates for clinical laboratory tests were based on market data from the "full spectrum of laboratories, including hospital, independent, and POLs." However, when CMS released its new proposed CLFS in 2017, more than 90% of the surveyed payment data used were derived from commercial laboratories, and only 1% were from hospital-based laboratories. This distribution of payment data is not representative of the laboratory testing market, where physician office and hospital laboratories account for approximately 43% of Medicare Part B spending for outpatient lab testing. It is likely that payment data disproportionately surveyed from independent labs (with higher volumes and lower cost margins) could skew prices downward and lead to deeper cuts to the CLFS than Congress had intended. While PAMA does place some limits on the reduction of the payment amount for an existing test as compared with the payment amount for the preceding year, the resulting data are projected to still lead to significant rate reductions in laboratory reimbursement, with Medicare annual cuts of 10% to most tests in the fee schedule each of the 3 initial years through 2020, and up to 15% annual cuts for the subsequent 3 years. After this 6-year phase-in period, there will be no limits to annual payment reductions.

Early examples of fraud and abuse

As CMS moved to a fixed fee schedule, laboratories, which had embraced automation and the technology for performing large numbers of tests at small incremental cost during the 1980s, saw that profits were limited by the fee schedule and began to reexamine business practices to increase their volume. In so doing, some laboratories created marketing tools to encourage physicians to order additional tests on their patients. These techniques were initially financially successful but were eventually deemed at least abusive, if not fraudulent.

It is instructive to examine the government's critical analysis of the practice of bundling tests together for ordering purposes and unbundling them for billing purposes. With reimbursement for tests limited by the fee schedule, some laboratories, especially for-profit laboratories, attempted to increase test volumes by assembling a package of tests directly to physicians at low costs. However, when the same packages of tests were offered

to Medicare beneficiaries, each test was individually charged to Medicare at higher prices. For example, a 20-test chemistry profile might be billed to the physician for $8.00, but the same test with an added TSH would be billed to the physician for $8.50. When the 20-test profile was billed to Medicare, Medicare paid $15.00. However, when a claim was submitted to Medicare for both the 20-test profile and a separate TSH, the reimbursement would have been $15.00 for the chemistry profile and an additional $23.00 for the TSH assay (total = $38.00). The position of the government was that the laboratory was deliberately misleading the physician to order tests that were unnecessary, or at least of minimal utility, but noting that the additional test (the TSH) was provided at only a very slight additional charge (the extra $0.50 above the charge for the chemistry profile), whereas Medicare was paying $23.00 more than the charge for the chemistry profile.

Government investigators viewed this tactic as a deliberate attempt to induce fraudulently the physician to order unnecessary tests. These charges were brought not under a theory of common law fraud that individuals and private corporations would utilize, but as a violation of federal statutes, most importantly the False Claims Act.

The False Claims Act

The False Claims Act, passed in 1863, was in response to war profiteers during the Civil War who sold inadequate or nonexistent war materials to the Union Army. The law provided for a high fine for each false claim that was submitted to the government for payment and a reward to the individual who informed the government of the fraudulent claim. Following the Civil War, the law fell into disuse, but it was revived in the 1980s in response to news stories of defense contractors selling items at inflated prices to the military during the Cold War, an era marked by reports of governmental waste such as the $435 hammer and the $640 toilet seat. Congress changed the False Claims Act by increasing the fine per false claim to as much as $10,000, providing the whistleblower who filed the charges against the defendant up to 30% of the total recovery, and requiring the defendant to pay the legal fees of the whistleblower. Because each Medicare claim was considered a separate false claim, it was possible for the total fine assessed against a large high-volume laboratory to reach into the hundreds of millions of dollars. In addition, any entity convicted of filing false claims could be barred from participating in Medicare, essentially ensuring the demise of that company. Laboratories who had engaged in these practices were willing to settle all charges for fines that ranged up to $325 million, provided they were not required to admit wrongdoing and, more importantly, provided they were not barred from participating in Medicare work in the future.

Inducements and kickbacks

The Social Security Act contains a provision that bars the payment of cash or the provision of any item of value when it is given for the purpose of inducing or having the effect of inducing healthcare providers or Medicare beneficiaries to order or use a Medicare benefit. This law applies to laboratories that induce a physician to use a specific laboratory's services. Examples include cash bribes as well as providing computers or fax machines. While computers may be permissible if they are for the exclusive purpose of delivery of test results, these same items can be considered an impermissible inducement if used by the physician for purposes unrelated to the lab test. More creative, but equally illegal, kickbacks include rental of office space for the storage of phlebotomy supplies and payment of consulting fees to client physicians. Under the antikickback law, it is equally criminal to offer, provide, request, or receive goods, services, or anything of value to induce the use of Medicare services; parties on both sides of such a transaction can be charged and punished. The government does not need to prove that there was specific intent to provide the inducement, only that the provider should have known that the item or service of value is likely to influence the beneficiary. This standard is met when a provider acts with deliberate ignorance or reckless disregard to the possibility of inducing a referral.

Less clear-cut is the billing practice when a physician refers both Medicare and non-Medicare testing to a laboratory. The laboratory may legally discount the non-Medicare testing but, to stay clearly within the law, the discount must be reasonable from a business point of view. Discounts may not vary in any way relating to the volume of Medicare tests sent and may never be below the actual cost of testing. To help guide providers faced with such decisions, both laboratories and physician practices may request an advisory opinion from the HHS OIG. The OIG's advisory opinions are published on its website (http://www.oig.hhs.gov), and although they have legal force only for the parties submitting the question, they are a valuable source of guidance and insight into the OIG's legal logic in evaluating potentially illegal activities.

Office of the Inspector General's Compliance Guidelines

The OIG has published compliance guidelines for many healthcare entities, and they have all followed the same organizational pattern. As such, the guidelines are general in nature and only occasionally give detailed guidance. In

addition to clinical laboratories, there are multiple other healthcare entities that have received OIG compliance program guidance (Table 15.10).

The guidelines that the OIG have published do not carry the force of law, but if they are carefully put into practice in a healthcare organization, such compliance can be offered as evidence of lack of criminal intent in the event of a prosecution for fraud. Committing Medicare fraud is intentionally submitting false information to the government or government contractor to receive money or a benefit. Submitting bills for tests not performed or on fictitious patients is an example of fraudulent activities. A common defense in the face of fraud charges is that the provider did not see anything wrong with their coding or billing practices, or their marketing or business relationships. In other words, whatever the fraudulent act, the provider will assert that they did not deliberately set out to defraud the government. An accused healthcare entity that can produce evidence that it did not knowingly seek to defraud the government has a potential defense against the government's case. Following the OIG's guidelines can be helpful in that effort to demonstrate a lack of fraudulent intent.

The OIG's guidelines for laboratories contain the following recommendations. (1) The laboratory should have written compliance procedures and policies. These should give direction to employees and there should be written notices to physician clients regarding Medicare policy and reimbursement. Standards of conduct should be defined, and billing practices should be clearly described. Adherence to the written standards of conduct should be an element of performance evaluations for staff and managers. (2) The laboratory should designate a compliance officer and, depending on its size and complexity, a compliance committee. The compliance officer should have sufficient independence to investigate any area of the laboratory and should report directly to senior management. The committee or individual should be involved in developing the standards of conduct described earlier. (3) Training and education should be conducted regularly. The laboratory should ensure that all employees, new as well as existing, are educated regarding its compliance policy and practices. (4) There must be effective lines of communication between the employees and the compliance officer. This includes a requirement to keep records of complaints, investigations, and corrective actions. (5) Compliance standards must be enforced through disciplinary practices that are publicized and understood by employees. This includes, specifically, holding corporate officers and managers accountable for enforcing the compliance policies. (6) There must be a process for auditing and monitoring the laboratory's practices. This includes coding and billing practices as well as sales and marketing programs. The auditing process must be conducted in sufficient detail that any significant increase in the billing for individual current procedural terminology (CPT) codes will come to the attention of the compliance officer, so they can consider whether this is the result of improper marketing efforts. (7) Corrective action must be undertaken if violations are detected. Misconduct must be reported to the government, and improper payments must be returned.

The largest settlements (amount of money including fines and reimbursement) paid by laboratories occurred during the mid and late 1990s, just before the issuance of the OIG's compliance guidelines. A review of these settlements makes it clear that upward of 80% of total settlement amounts were payments involving for-profit laboratories. This may be an expected consequence of a for-profit business using creative and sometimes illegal means to enhance revenue.

Not surprisingly, the OIG has focused its guidance on the wrongdoing of these laboratories. However, not-for-profit laboratories, such as those based in not-for-profit hospitals, have entirely different organizational goals and correspondingly different paths to achieve those goals. As a result, in the past, nonprofit entities have enjoyed less scrutiny. However, the need to sustain and grow healthy profit margins in any organization can incentivize even a nonprofit to run into compliance issues. Therefore while the OIG's compliance guidelines may be less applicable for a laboratory operating in a not-for-profit environment, the guidelines should remain a reference point. The benefit of these guidelines and of the OIG's advisory opinions is that they help laboratory managers understand the arms-length relationship that the OIG expects laboratories to maintain with their clients. Regardless of profit incentive, organization winning business through improved service and price is perfectly acceptable, but through any other kind of benefit is strictly forbidden.

Direct billing

The Social Security Act, where it states requirements for reimbursement for services under Medicare, contains a provision that limits reimbursement to only the healthcare provider who actually provided the service. This provision, referred to as the "direct billing" requirement, was considered necessary to prevent a healthcare provider (e.g., a physician practice) from shopping around different laboratories to obtain the lowest priced services, below the fee schedule amount, and then profiting from the difference between what Medicare reimbursed the provider for the tests and what the physician practice actually paid to the laboratory for the tests that were performed. Quality and abuse concern arise from such arrangements. Laboratories and physician offices may compromise appropriate handling and testing protocols to make up for

TABLE 15.10 Office of the Inspector General compliance program guidance for healthcare entities.

Year	Healthcare segment	Reference
1998	Clinical laboratories	63 Fed. Reg. 45076
1998	Hospitals	63 Fed. Reg. 8987
1998	Home health agencies	63 Fed. Reg. 42410
1998	Third-party medical billing companies	63 Fed. Reg. 70138
1999	Durable medical equipment, prosthetics, orthotics, and supply industry	64 Fed. Reg. 36368
1999	Hospices	64 Fed. Reg. 54031
1999	Medicare + Choice organization	64 Fed. Reg. 61893
2000	Nursing facilities	65 Fed. Reg. 14289
2000	Individual and small group physician practices	65 Fed. Reg. 59434
2003	Ambulance suppliers	68 Fed. Reg. 14245
2003	Pharmaceutical manufacturers	68 Fed. Reg. 23731
2005	Recipients of public health service research awards	70 Fed. Reg. 71312

the lost margin for the laboratory. It also creates an incentive for the provider to order as many tests as possible, because the physician practice would enjoy a profit from each test ordered and performed. The direct billing requirement, which mandates that the laboratory be reimbursed directly by Medicare, eliminates the incentive to the physician practice of ordering excess tests. Unfortunately, novel creative business practices have emerged to circumvent the direct billing requirement.

Stark amendments

The direct billing rule resulted in the creation of business entities that could be paid for performing tests. Ordering physicians became owners or investors in contrived laboratories so they could profit from the tests that they ordered. In these "shell" or "sham" laboratories (owned by physicians or physician groups), all the actual testing would be referred to an outside laboratory; usually, it was the outside lab that would set up the shell laboratory entity. The shell laboratory would bill and be reimbursed by Medicare, and profits would be distributed back to the owning physicians or physician practices, sometimes in direct proportion to the number of tests referred. In some cases, the shell laboratory was little more than a location used to redirect all specimens to the laboratory that set up the shell arrangement and performed the actual testing. A well-publicized lawsuit involving a California shell laboratory (*Hanlester Network v. Shalala*) resulted in a victory for the defendant laboratory and physicians involved, unfortunately based mostly on legal technicalities.

To eliminate this practice, laws were introduced by Congressman Pete Stark, eventually becoming known as the Stark Amendments, which strictly prohibited a physician from sending laboratory tests to a laboratory in which s/he or a family member had a financial interest. The law was later expanded to include other entities, such as radiology or physical therapy facilities, and this expansion became known as Stark II. Both can be considered two parts of the same concept—prohibition of physicians referring a Medicare patient to a laboratory or other designated health services if the physician or an immediate family member of the physician has a direct or indirect financial interest in the entity providing such services. A financial interest includes an ownership interest or a compensation arrangement. There are also reporting requirements under the Stark Amendments. Entities furnishing items or services for which payment may be made under the Medicare or Medicaid programs must submit information to CMS concerning their financial relationships.

Key financial compliance considerations

The bottom-line is that inducing a physician or a patient to order or receive laboratory tests, or any other healthcare service, for reasons not related to medical necessity is strictly prohibited. The OIG issued an alert in 1994 that specifically forbade the following practices:

- Advertisements that state, "Medicare accepted as payment in full," "insurance accepted as payment in full," or "no out-of-pocket expense."

- Advertisements that promise that "discounts" will be given to Medicare beneficiaries.
- Routine use of "financial hardship" forms that state that the beneficiary is unable to pay the coinsurance/deductible.
- Collection of copayments and deductibles only where the beneficiary has Medicare supplemental insurance coverage.
- Charges to Medicare beneficiaries that are higher than those made to other persons for similar services and items.
- Failure to collect copayments or deductibles for a specific group of Medicare patients for reasons unrelated to inability to pay.
- Programs that cover copayments or deductibles only for items or services provided by the entity offering the insurance.

It is important for laboratorians to develop an appreciation for the guidelines expressed by the OIG and others involved in regulation of government reimbursement for healthcare services. Practices that seem, on the surface, to be acceptable may be strictly prohibited, and developing sensitivity to suspect practices is the best way to remain compliant.

Food and Drug Administration oversight of laboratory developed tests

Over the past two decades, there has been a debate within US healthcare regarding federal regulation of laboratory developed tests (LDTs) [6]. An LDT is defined by the FDA as "an IVD that is intended for clinical use and design, manufactured and used within a single laboratory." These tests are currently regulated under CLIA '88 as high-complexity tests. CMS estimates that approximately 11,000–12,000 laboratories have the needed certification to perform this level of testing. Therefore only about 5% of the more than 250,000 laboratories in the United States can perform LDTs—most often in academic medical centers, large commercial laboratories, and hospitals.

The FDA asserts that it has statutory authority to regulate LDTs under the Medical Device Act of 1976, but previously deferred to CMS under its "enforcement discretion." However, the growing complexity of LDTs has forced the agency to reevaluate its position. The FDA released two draft guidance documents outlining how it proposes to regulate LDTs. Many laboratories argue that complying with the proposed new requirements would result in the cessation of performing LDTs. The FDA proposal to expand its oversight to LDTs has divided the broader healthcare community in two separate camps. One side argues that enhanced LDT oversight should be carried out under the current CLIA regulatory structure. The other side argues that the FDA should expand its oversight. Legislators have focused more on the larger issues of process and what, if any, role the FDA should play in LDT oversight. The FDA's role in LDT oversight is not yet determined.

State regulations

State law traditionally revolves around issues such as interpretation of contracts, general health and welfare, and public safety. Insurance law is another area specific to state law that affects healthcare organizations, particularly regarding scope of coverage and policy determinations, and occasionally, that may affect, for example, whether a patient's out of state reference lab testing is covered under a policy. However, state law usually does not have a direct impact on laboratory oversight. Many states, however, have laws and regulations that affect all businesses such as employee protections and insurance regulations that apply to healthcare businesses just as any other business. One way to become familiar with general business regulations unique to a state is to obtain business-oriented publications from a chamber of commerce or similar organizations that serve the business needs of the state. Unique laboratory regulations can be obtained from state laboratory or hospital organizations, from the state public health department, libraries, and of course, Internet sites.

Negligence

The last area of regulation, tort law, is not a regulation at all in the sense that it requires compliance to a set of rules outlined by a governmental agency, but its presence and effect is as important as any law. It is how private individuals can seek a civil remedy for an injury outside of specific laws and regulations. Medical negligence is also a form of tort law. Under common law, a person can bring a lawsuit seeking compensation for injury—this is the basis of medical malpractice cases. The essential elements to prove compensable damages are establishing duty or obligation among the parties of the suit, a breach of that duty, and a resulting injury directly related to the breach. In the case of a laboratory, a laboratory performing testing has a duty to perform the test in a reasonably careful manner. If, because of a violation of that duty, it results in some error and if a patient suffers from some injury as a result of the error, the patient can claim damages. While the laboratory as an independent entity or part of a larger organization can be liable for patient injury under tort theory, the laboratory medical director can be held

professionally responsible for a breach of duty by abrogating the responsibilities outlined in the CLIA regulations for the laboratory director. However, in terms of professional liability, laboratories and their medical directors enjoy relatively low rates of negligence or malpractice actions. Nonetheless, because a laboratory error has the potential to cause injury, those responsible for testing need to be aware of the risks and to minimize them whenever possible.

Acknowledgments

The authors gratefully acknowledge the original contribution by Robert L. Murray and Charles Root, upon which portions of this chapter are based. The information in this chapter provides general information but should not be considered a definitive reference for regulatory compliance.

References

[1] Department of Health and Human Services, Office of Inspector General. Publication of OIG compliance program guidance for clinical laboratories, Fed. Regist., 63, 1998, pp. 45076–45087.

[2] Top 10 Conditions in the Nation – CMS Surveys (2018). <https://www.cms.gov/Regulations-and-Guidance/Legislation/CLIA/>.

[3] Clinical Laboratory Improvement Amendments of 1988. PL 100578, Sec. 353(d)(3)(B–C).

[4] Food and Drug Administration, Draft Guidance for Industry and FDA Staff: Recommendations for Clinical Laboratory Improvement Amendments of 1988 (CLIA) Waiver Applications, Department of Health and Human Services, Washington, DC, 2005.

[5] D.M. Wolman, A.L. Kalfoglou, L. LeRoy (Eds.), Medicare Laboratory Payment Policy, Now and In the Future, National Academy Press, Washington, DC, 2000.

[6] J.R. Genzen, J.S. Mohlman, J.L. Lynch, M.W. Squires, R.L. Weiss, Laboratory-developed tests: a legislative regulatory review, Clin. Chem. 63 (10) (2017) 1575–1584.

Self-assessment questions

1. The federal regulation that most directly affects clinical laboratories is:
   - **a.** OSHA regulations
   - **b.** CLIA
   - **c.** HIPAA
   - **d.** Stark amendments
2. Compliance with the HIPAA privacy standard includes the following:
   - **a.** control of paper containing PHI
   - **b.** limiting inclusion of PHI on reports and worksheets
   - **c.** recording of all every instance of release of PHI that is not specifically related to treatment, operations, and payment
   - **d.** shredding of documents containing PHI
   - **e.** all of the above
3. What is the legislation that prohibits a physician from sending laboratory tests to an entity in which they have a financial interest?
   - **a.** Stark amendments
   - **b.** CLIA '88
   - **c.** HIPAA
   - **d.** The False Claims Act
4. What role can a certified clinical chemist or clinical pathologist fill in a clinical laboratory providing moderate- or high-complexity testing?
   - **a.** Director
   - **b.** Clinical consultant
   - **c.** Testing personnel
   - **d.** a and b
   - **e.** All of the above
5. Which of the following tests are "waived" under CLIA?
   - **a.** Urine dipsticks
   - **b.** Erythrocyte sedimentation rate
   - **c.** Rapid strep test
   - **d.** a and b
   - **e.** All of the above
6. Which Federal agencies are tasked with administration of CLIA?
   - **a.** CLIAC
   - **b.** CMS
   - **c.** FDA
   - **d.** CAP
   - **e.** All of the above
7. Which of the following organizations is not considered "deemed" by CMS for the purposes of clinical laboratory accreditation?
   - **a.** Joint Commission
   - **b.** American Osteopathic Association
   - **c.** American College of Medical Genetics and Genomics
   - **d.** College of American Pathologists

Answers

1. b
2. e
3. a
4. e
5. e
6. b
7. c

Chapter 16

Evidence-based laboratory medicine

Alex C. Chin[1], Robert H. Christenson[2], and Stephen E. Kahn[3]

[1]*Alberta Precision Laboratories and Department of Pathology and Laboratory Medicine, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada,* [2]*Department of Pathology, University of Maryland School of Medicine, Baltimore, MD, United States,* [3]*Department of Pathology, Loyola University Medical Center, Maywood, IL, United States*

Learning objectives

After reading this chapter the participant will be able to:

- Explain each of the basic evidence-based laboratory elements of asking the clinical question, acquiring the evidence, appraising and analyzing the evidence, applying the findings to the clinical problem, and how to assess or audit the process.
- Define the PICOTS strategy for formulating a clinical question. Provide two examples of clinical questions developed in the PICOTS format.
- List five different types of investigative studies listing a strength or a limitation for each type of study.
- Sketch or write expressions for the ROC curve, likelihood ratios, odds ratios, and diagnostic sensitivity, and specificity.
- List and describe five resources and/or tools that are useful in the application of EBLM for the purpose of translating the evidence into changes in the practice of laboratory medicine.

Introduction

In 2001 the Institute of Medicine (IOM) reported that there was a massive chasm between what is known to be good health care in the United States and the health care that patients actually receive. To help address this mismatch, the IOM stated that health care should strive to be safe, timely, efficient, effective, equitable, and personalized. Evidence-based medicine consists of tools that can help meet these goals and assist in bridging the healthcare gap.

Evidence-based medicine (EBM) was initially described as the "use of the best data available to generate unbiased informed decisions and treatment of patients." Although by this definition, EBM practice may appear to focus on the science of medicine, "best data" must also include the clinical wisdom accumulated over years of experience combined with clinical intuition and understanding of healing. Further, once available alternatives and recommendations have been discussed, patient preferences and expectations must be considered a most important component in decisions regarding treatment and management. A more evolved definition of clinical EBM is "the integration of the best research evidence with clinical expertise and patient values." Evidence-based laboratory medicine (EBLM) differs in that it has components both before and after the clinician–patient discussion on how to manage disease. Before this discussion, EBLM involves the use of the best research evidence and laboratory medicine expertise to aid in disease diagnosis and stratifying risk. After patient management decisions, laboratory evidence and expertise are used to monitor the effects and outcomes of such decisions.

Traditionally, basic science, pathophysiology, and pharmacology were the foundation of medical and laboratory training used to decide if an action makes sense. During training, learning at the bench and bedside was largely from the experience of master laboratory professionals and clinicians. Once in practice, individuals learned and remained updated through continuing education lectures and seminars from thought leaders. Moreover, clinical experience, which stressed the successes, outcomes, and adverse events in our own experience, guided future behavior. This traditional approach has been problematic, however, because professional judgment improves over time, up-to-date knowledge of the best, most current evidence may become obsolete. There have been numerous studies that demonstrate suboptimal adoption of proven therapies and management strategies, because the best evidence must frequently be revised, updated, disseminated, adopted, and adhered to. Information from expert advice and textbooks necessarily lags behind the current medical and scientific literature. On the other hand, because the pace of scientific discovery and production of peer and nonpeer-reviewed literature is increasingly rapid, determining which information

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00016-8

represents valid and relevant scientific evidence is complicated and time-consuming. EBLM can facilitate the task of laboratory professionals, clinicians, managers, and policymakers faced with complex decision-making when the information is required in a real-time environment. In addition, EBLM can aid in the task of utilization review both at the Laboratory Medicine and Institutional levels by providing evidence-based guidance for translating, tailoring, and implementing effective testing systems and strategies. Use of EBLM can enhance the value of laboratory medicine by optimizing health outcomes and efficient resource deployment in the preanalytical, analytical, and postanalytical phases of testing.

Elements of evidence-based laboratory medicine

The clinical question

Fig. 16.1 outlines an example of a cyclic process utilized for practicing EBLM. This example was specifically developed for the Center for Disease Control's (CDC's) Laboratory Medicine Best Practices program. It is known as the CDC-LMBP A-6 Systematic Steps or the A-6 Cycle [1,2]. Additional information on projects and activities conducted under the auspices of the CDC-LMBP can be found at https://www.cdc.gov/labbestpractices/index.html. As illustrated in Fig. 16.1, the first step of the EBLM cycle is to frame focused question(s) that are to be answered by the evidence review. Once a focused clinical question has been identified, the evidence is acquired, appraised, analyzed, and applied, and the application(s) are audited and assessed. Following these six steps then provides the option to cycle back to a need for framing a new focused question to be answered. In this way, the process evolves into another A6 cycle addressing a new, further refined or related evidence review.

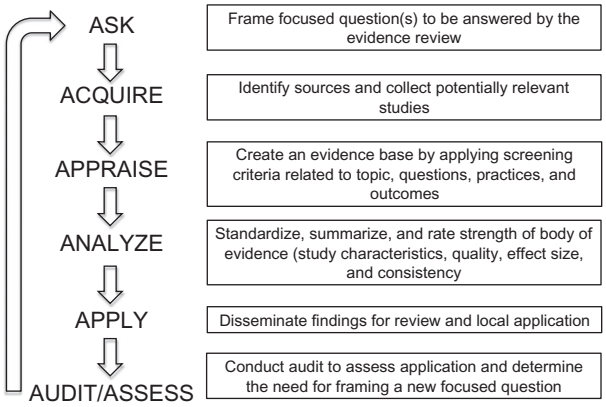


FIGURE 16.1 The A6 cycle is shown as an example of the elements of the evidence-based laboratory medicine process. Use of the A6 cycle to review systematically the literature is one aspect of the Centers for Disease Control Laboratory Medicine Best Practices program.

Formulation of a clearly stated and clinically relevant question is an important first step necessary for effective practice of EBLM. Key reasons why physicians and other caregivers order laboratory medicine tests on individual patients are to screen for a variety of conditions, rule in or rule out a diagnosis, monitor a response to therapy, assess prognosis for long-term outcomes, and guide therapy or interventions. A widespread and convenient method for initially formulating focused, clearly stated, and clinically relevant questions uses a strategy represented by the acronym "PICOTS" to explicitly state the framework for the question being asked and to initiate the review of the evidence. As shown in Table 16.1, PICOTS stands for *p*atient (clinical situation), *i*ntervention (or assay), *c*omparison (reference or control population), *o*utcome(s), *t*iming, and *s*etting(s). The example provided in Table 16.1 for a key question focuses on the use of procalcitonin (PCT) measurement in addition to clinical criteria for infection compared with clinical criteria for infection alone for the purposes of guiding antibiotic therapy and affecting outcomes in patients. This question became the basis for a comparative effectiveness review (CER) conducted by the Agency for Healthcare Research and Quality [3]. The elements of PICOTS for this question are provided in Table 16.1.

As the landscape of healthcare delivery continues to transform rapidly across the world, the need to conduct EBLM studies focusing on clinical questions considerably more complex than the diagnostic accuracy of laboratory tests has become the norm. The clinical question posed for the PICOTS elements in Table 16.1 is a real example. The utility of PCT to guide antibiotic therapy in selected patient populations extends beyond the analytical quality and diagnostic accuracy of PCT tests. The clinical question was structured to compare whether patients managed with a diagnostic intervention (use of PCT testing) coupled with other clinical criteria fared better than patients who were managed with clinical criteria alone. As this evidence-based review reported, the analytic framework ultimately focused on the value of the intervention in affecting intermediate outcomes as well as health outcomes.

Finding the evidence

The PICOTS strategy is useful not only for focusing question formulation but also for finding evidence that addresses the question. Keywords derived from the PICOTS structure can be used to search the literature directly in the context of the clinical question. Perhaps, the most useful resource for finding evidence is PubMed, a service provided at no charge by the United States

TABLE 16.1 Components and example of structured PICOTS[a] relating to the key question[b].

PICOTS	Questions to consider	Specific of each element of PICOTS
P: Patients or patient population(s)	What are patients' characteristics, symptoms, and demographics? Consider the condition(s) that may be present	Adult and pediatric patients with known or suspected local or systemic infection, including critically ill patients with sepsis syndromes or ventilator-associated pneumonia, adults with respiratory tract infections, neonates with sepsis, children with fever of unknown source, and postoperative patients at risk of infection
I: Intervention	What laboratory medicine test is being considered? Are other criteria being considered?	Initiation, discontinuation, or intensification of antibiotic therapy guided by procalcitonin plus clinical criteria for infection
C: Comparator (or control)	What is the reference or "Gold standard" for the intervention?	Initiation, discontinuation, or intensification of antibiotic therapy guided with procalcitonin compared to clinical criteria alone.
O: Outcome	What is (or what are) the endpoint(s) of interest?	Antibiotic use (duration of antibiotic therapy, prescription rate, and total antibiotic exposure), mortality, morbidity (length of stay and severity of illness score), total cost, and adverse events of antibiotic therapy (persistent or recurrent infection, and antibiotic resistance)
T: Timing	What is the target time frame for studies to have been conducted?	3 months
S: Settings	What are the clinical settings for studies to have been conducted?	ICUs (medical and surgical), inpatient acute care hospitals, emergency departments, and outpatient clinics

- Intermediate outcomes, such as initiation, discontinuation, or change of antibiotic therapy; antibiotic use; and length of stay?
- Health outcomes, such as morbidity, mortality, function, quality of life, and adverse events of antibiotic therapy (persistent or recurrent infection, and antibiotic resistance)?

[a]PICOTS, *Patient, intervention or test, comparator or control, outcome, timing, and settings*
[a]*In selected populations of patients with suspected local or systemic infection, what are the effects of using procalcitonin measurement plus clinical criteria for infection to guide initiation, discontinuation, or a change in antibiotic therapy when compared with clinical criteria for infection alone on:*
Source: Adapted from Procalcitonin-guided antibiotic therapy. Vol. 2012.

government. The Internet address for PubMed is http://www.ncbi.nlm.nih.gov/pubmed/. The Boolean logic terms AND, OR, and NOT are used to combine keywords, authors, and terms either to limit or broaden the search strategy. For example, to search the PICOTS question stated above, one could begin by entering the following terms: PCT, infection, and antibiotic therapy.

Resources for searches on diagnostic tests

Although there are numerous databases available electronically, it is important to recognize that the Internet is a dynamic tool. Information can go out of date, links may change, new databases may appear, and so on. Table 16.2 summarizes resources that provide useful information when conducting searches for relevant evidence. Some databases listed here are free and some are provided at a cost; also, some databases are from a single source, whereas others are combined resources. Consulting experts in library science as well as the particular field of interest can assist in finding potentially useful information. Further, unpublished studies may be located by examining bibliographies of identified articles and literature from governmental institutions.

Approaches to evidence-based laboratory medicine—types of analyses

As mentioned earlier, EBLM has components before and after the clinician–patient interaction and is a separate branch of EBM pertaining to the use of laboratory tests to improve patient outcomes. Diagnostic tests are frequently released for clinical use without adequate assessment of clinical effectiveness. Furthermore, the rising costs of health care and the ease of quantifying laboratory processes make it even more imperative to ensure that resources are used most efficiently with the best clinical evidence. Given the importance of health outcomes, the availability of tools and resources devoted to evidence-based practice and EBLM has steadily evolved from opinion-based narrative reviews to objective, structured systematic reviews (SRs). Before searching for evidence, it should be noted that there is a hierarchy of evidence that needs to be considered (Table 16.3). Randomized controlled trial (RCT) SRs and meta-analyses are at the top, while expert opinions (narrative reviews) and case reports are at the bottom [5].

SRs are structured, usually more resource-intensive, and use defined methods and strategies to locate, evaluate,

TABLE 16.2 Selected resources for searches involving laboratory medicine tests.

Source	Description	URL
Resources provided at no cost		
PubMed	Free and allows use of validated filters for selecting papers on diagnosis, etiology, prognosis, and therapy	http://www.pubmed.com
Lab Tests Online	Health information web resource produced by AACC and designed by laboratory professionals to help patients and caregivers understand the clinical utility of laboratory tests	http://www.labtestsonline.org
Academic discussion groups	After searching for evidence using the above resources, one could ask members of the group for references or discussion on a specific topic	For instructions on how to join these and other lists: • http://jiscmail.ac.uk/ • aacc-eblm@aacclists.org (Listserv)
WebMD MedicineNet Medscape eMedicine	The WebMD health professional network Includes up to date information across a broad range of specialty areas from experts in a user-friendly format	http://www.webmd.com http://www.medicinenet.com http://www.medscape.com http://www.emedicine.medscape.com
Google Google scholar Yahoo	Search engines for a variety of topics and scholarly literature, but will direct the user to both peer- and nonpeer-reviewed information	http://www.google.com http://scholar.google.com http://www.yahoo.com
Resources requiring user fees for full access		
The Cochrane Library British Medical Journal Clinical evidence	Contains 7 databases; the DARE database is the most relevant Summarizes information from therapeutic trials and will include information on diagnostic procedure	http://www.cochrane.org http://www.clinicalevidence.com
Annals of Internal Medicine—ACP Journal Club	Critically appraises well-performed and valid studies and has a focus on diagnostic studies In May 2008, *ACP Journal Club* ceased being a stand-alone journal and now exists as a part of *Annals of Internal Medicine* under "Journal Club"	• Archives up to and including April 2008: http://www.acpjc.org • May 2008 to current: http://annals.org/journalclub.aspx
TRIP database	Clinical search engine that provides evidence-based summaries with comments, also on diagnostic issues	http://www.tripdatabase.com
OVID	A commercial database with access to evidence-based resource material; gives access to EMBASE, one of the major databases needed for performing searches when doing systematic reviews	http://www.ovid.com/index.cfm
Clinical Chemistry Journal of Applied Laboratory Medicine	Includes sections on EBLM and test utilization	http://www.clinchem.org https://www.aacc.org/publications/the-journal-of-applied-laboratory-medicine
Up-To-Date	An evidence-based, clinical decision support resource including evidence-based recommendations to improve patient care and quality	http://www.uptodate.com

ACP, American College of Physicians; *EBLM*, Evidence-Based Laboratory Medicine Committee.

and compile scientific evidence that addresses focused questions about healthcare issues of diagnosis, screening, treatment, monitoring, and preventive services. SRs are frequently prepared by investigators who are not necessarily experts in the field, using a systematic approach incorporating all the elements listed in Table 16.4 as recommended by the IOM to not only minimize bias and random error, so a "minimum level of objectivity, transparency, and scientific rigor is met" [6]. Indeed, the importance of standards are readily apparent when SRs focus on especially high impact questions that affect a broad population while using public resources. This ensures that the quality of SRs is maintained while being scientifically valid, transparent, and reproducible. Proper SR addresses the key questions by limiting the effects of bias, integrating relevant information from available evidence, distilling data to manageable levels, and identifying limitations in methodology. The ultimate goal of SR is to present information that clarifies what is known from what is speculated, identifies sources of

TABLE 16.3 Types of studies and their respective key domains.

Study	Key domains covered	Strengths	Weaknesses	Strength of evidence
Systematic reviews • Meta-analysis	Study question, search strategy, inclusion and exclusion criteria, data extraction, funding or sponsorship	• Strongest form of evidence • Comprehensive summary of multiple studies to answer a research question • Meta-analyses use statistics to evaluate the effectiveness particularly when there are inconsistencies between studies and different sample sizes • Standardization of methodology	• Publication bias (negative results are usually not published) • Low-powered studies may affect the meta-analysis in a negative manner • Potential heterogeneity between studies may lead to inappropriate conclusions	
Randomized controlled trial	Study population, randomization, blinding, interventions, outcomes, statistical analysis, funding, or sponsorship	• Most reliable form of scientific evidence for clinical research • Avoids bias and uses concurrent control group • Premature discontinuation of the study due to significant findings or perceived disadvantages in control subjects, but this will lead to incomplete data collection	• Results may not be reproducible between settings such as patient characteristics, study procedures, outcome measures, and incomplete reporting adverse events • Very expensive and time-consuming • Conflict of interest from commercial sponsors	

Observational studies
• Cohort
• Case-control
• Cross sectional
• Case report
	Comparability of subjects, exposure or intervention, outcome measurement, statistical analysis, funding, or sponsorship	• Cohort studies:
		  • Can help determine risk factors for disease
		  • Longitudinal study to track incidence of disease over time
		• Case-control:
		  • Relatively inexpensive
		  • Shorter than cohort studies
		• Cross sectional:
		  • Very inexpensive
		  • Can provide valuable epidemiological data on whole populations
		• Case reports:
		  • Provide resolutions to rare diseases or conditions and unusual presentation to common diseases
			• All observational studies indicate causal associations, but not true causality
			• Cohort studies:
			  • Expensive to conduct
			  • Prone to loss of active participation
			  • Long timeline delays the collection of data and conclusions	

Diagnostic test studies	Study population, adequate description of test, appropriate reference standard, blinded comparison of test and reference, and avoidance of verification bias	• Specifically focuses on a diagnostic test or testing strategies	• Reference method or true Gold standard may not exist or is not practical to be used as a comparator • Usually only focus on accuracy and cannot link to outcomes	

Source: Adapted from (AHRQ) AfHRaQ. Systems to rate the strength of scientific evidence. In: Services USDOHaH, Agency for Healthcare Research and Quality (AHRQ), Rockville, MD, 2002 [4].

TABLE 16.4 Institute of Medicine Standards for systematic reviews.

Standards for initiating a systematic review

1. Establish a team with appropriate expertise and experience to conduct the systematic review
2. Manage bias and COI of the team conducting the systematic review
3. Ensure user and stakeholder input as the review is designed and conducted
4. Manage bias and COI for individuals providing input into the systematic review
5. Formulate the topic for the systematic review
6. Develop a systematic review protocol
7. Submit the protocol for peer review
8. Make the final protocol publicly available, and add any amendments to the protocol in a timely fashion

Standards for finding and assessing individual studies

1. Conduct a comprehensive systematic search for evidence
2. Take action to address potentially biased reporting of research results
3. Screen and select studies
4. Document the search
5. Manage data collection
6. Critically appraise each study

Standards for synthesizing the body of evidence

1. Use a prespecified method to evaluate the body of evidence
2. Conduct a qualitative synthesis
3. Decide if, in addition to a qualitative analysis, the systematic review will include a quantitative analysis (meta-analysis)
4. If conducting a meta-analysis, then do the following:
   a. Use expert methodologists to develop, execute, and peer review the meta-analyses
   b. Address the heterogeneity among study effects
   c. Accompany all estimates with measures of statistical uncertainty
   d. Assess the sensitivity of conclusions to changes in the protocol, assumptions, and study selection (sensitivity analysis)

Standards for reporting systematic reviews

1. Prepare final report using a structured format
2. Peer review the draft report
3. Publish the final report in a manner that ensures free public access

COI, Conflict of interest.

Source: Adapted from Finding What Works in Healthcare - Standards for Systematic Reviews, National Academies Press, Washington, DC, 2011.

TABLE 16.5 Types of biases to consider when reviewing the evidence.

Time-lag bias	Often long time required for a study to appear in print
Publication bias	Publication of the same data multiple times
Outcome reporting bias	Some important outcomes are frequently left unreported in publications
Language bias	Virtually all studies are published in the English language, and thus are more likely to be cited and used in systematic reviews
Indexing bias	Higher probability that positive results, published in leading journals, are cited and indexed more frequently in easily accessible search tools (e.g., Medline) compared with negative studies
Population bias	All participants are not equally balanced or objectively represented
Verification bias	Results are not evaluated with the reference or gold standard

inconsistency, aids in development of future research questions, and instills confidence that the review outcome represents thoughtful interpretation of the best evidence. Nevertheless, SRs can be subject to bias, whereby negative findings are less likely to be reported in the literature or inclusion of nonpeer reviewed unpublished studies may lead to overly optimistic conclusions. As a generally accepted practice, it is recommended that the effects of unpublished studies and data on the overall conclusion be evaluated by performing separate meta-analyses with and without the unpublished studies to evaluate consistency. Other biases are summarized in Table 16.5.

SRs incorporate summaries of data whenever possible. Perhaps, the most important of these summaries is termed a meta-analysis, which entails statistical analysis of results from similar but independent primary studies, meeting specific and clearly stated aims and criteria to produce a single overall estimate and uncertainty. The distinction between the SRs and the meta-analysis is important, although both involve the summarization of data. In principle, systematically reviewing a body of data is always appropriate and desirable; however, employment of statistical methods to summarize these results with a meta-analysis can produce a stronger and more objective conclusion. Sometimes, the statistical meta-analysis of pooled data from separate primary studies is not appropriate in all cases, because there is always some measure of heterogeneity with regard to subject inclusion, assay use, and other variables that can yield inappropriate or misleading results. Nevertheless, meta-analysis is performed to summarize the results of multiple studies, and expresses the most accurate reflection of the true effect or risk of a treatment or a test's true performance, addressing a well-defined research question in a homogeneous population of patients. Meta-analysis is considered by many experts to be a scientific endeavor and is used whenever appropriate as a method of summarizing and analyzing data during the performance of an SR. Therapeutic and device trials frequently use meta-analysis to evaluate and better define the effects of interventions. This technique is particularly powerful when there are conflicting results among a set of studies, some of which have large sample sizes. In such cases, analysis of a set of studies that investigates differences in research design may lead to explanations of the conflicting results. Meta-analysis is also helpful when there are inconsistently positive findings in a set of studies, none of which has large sample sizes. In this case, combining the results of several studies may give sufficient sample size and power to help answer a specified question.

Similarly, meta-analyses of test accuracy studies involve the computation of a weighted average of individual study results, which can improve the precision of estimates of test accuracy and reveal differences between the studies that affect test performance. The results are frequently summarized by "Forest plots." An example Forest plot is displayed in Fig. 16.2 for providing data to address the question: Are barcoding practices effective at reducing patient specimen and laboratory testing identification errors? [7]. This example allows the reader to examine the heterogeneity of data from nine studies examining this question. The overall summary effect in the lower portion of Fig. 16.2 provides a better notion of the actual odds ratio and uncertainty to be expected. However, it should be noted that individual biases associated with each study may be compounded and lead to inappropriate conclusions [8].

RCTs are studies in which at least one "test" treatment, intervention, or diagnostic strategy is compared with a control group that receives placebo, a different, or no intervention, or a different diagnostic strategy. RCTs typically enroll patients fitting prespecified criteria; subjects are assigned to receive either the test or control arms of the study by a random process. Patients assigned to

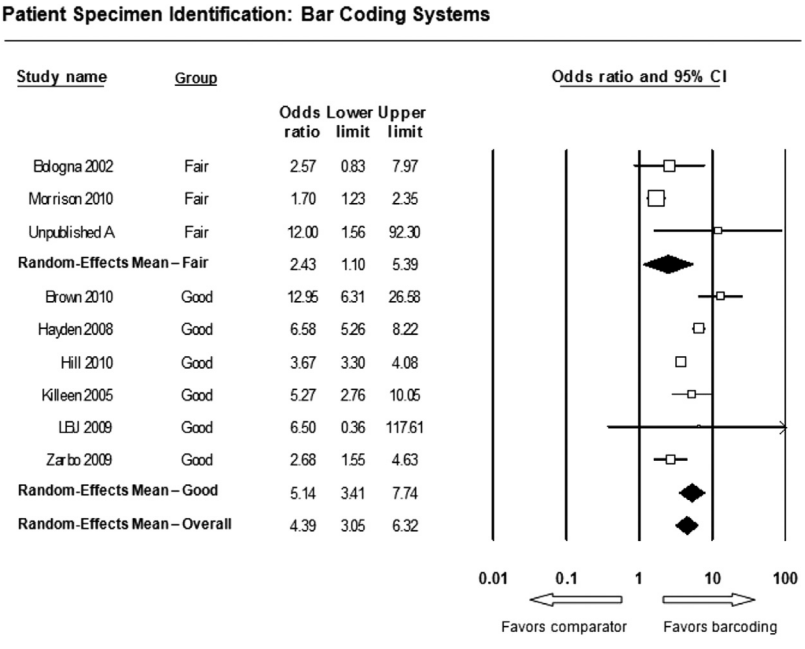


FIGURE 16.2 Forest plot showing the meta-analysis effect size results for barcoding systems practice over the nonbarcoding practice for improving identification error rates indicating a consistent and statistically significant effect. The mean odds ratio for an individual study is indicated to the far left, with the box size is proportional to the study sample size. The endpoints of the lines on the left and right sides of the box represent the lower and upper limits, respectively, of the study odds ratio's 95% confidence interval, with the numerical values provided to the left. The bottom line represents the overall summary effect (or grand mean) for all the studies in the body of evidence along with its confidence interval. In addition, meta-analysis results were tabulated separately for two subgroups using the two study quality ratings "Fair" and "Good." *Source: Reprinted with permission from Clinical Biochemistry and from Elsevier Inc. Copyright 2012 [7].*

either the test or control arms are concurrently enrolled, and there is complete follow-up of both. RCTs are a reliable method for determining whether a treatment, intervention, or diagnostic strategy is different from that in controls. The RCT design is superior for comparing interventions, because it avoids many biases and uses a concurrent comparison group (controls) to compensate for any trends over time. Characteristics of a well-designed RCT include: (1) an appropriate randomization procedure for unbiased allocation of patients to the test or control arm; (2) an appropriately sized sample to provide the statistical power to allow meaningful comparison; (3) an intention-to-treat analysis (results are based on initial treatment assignment and not on the treatment eventually received) that includes all individuals randomized to the test group; and (4) statistical results that include a specific *P*-value with confidence interval. Provided the patient enrollment criteria and aims are similar, data from RCTs may be pooled for meta-analyses to generate a more accurate and precise estimate of the effect of the treatment, intervention, or diagnostic strategy. Indeed, eligibility criteria for patient enrollment such as sex and race of each RCT should be considered when conducting an SR and meta-analysis as they can affect the interpretation.

Observational studies are not designed to alter or influence patient or practitioner behavior. Subjects are monitored either prospectively or retrospectively during their episode of care and no effort is made to change treatment, intervention, or testing strategy. Observational studies provide useful information about care patterns and patient preferences. They are also useful for generating hypotheses for future studies. However, these studies are frequently susceptible to patient-selection bias and may lead to inaccurate conclusions about the comparative benefit of various treatments, interventions, or testing strategies. This is because observational studies may be comparing patient groups with different prognoses due to temporal trends in disease characteristics, diagnostic methods, or supportive care. Observational studies can be described as a cohort, case-control, cross-sectional, or in the most simple form, a case report.

Cohort studies are longitudinal and follow a group (cohort) of people who share a common characteristic such as a disease state, treatment of a drug, or exposure to an environmental agent, and are compared with a group taken from the cohort general population. Cohort studies can be conducted retrospectively or prospectively by studying the group before or after the intervention of interest (e.g., disease state, treatment, and exposure), respectively. Ideally, the cohort is identified before the disease in question in order to determine risk factors (prospective cohort study). Conclusions from cohort studies cannot refer to cause and effect, but can only draw associations from the results. The advantage of cohort studies is that longitudinal observation allows collection of data at regular intervals and can help determine associations and risk factors over the natural history of the disease. By the same token, the disadvantages of cohort studies are due to their long-term framework and associated costs to conduct the research. Case control studies are conducted where the disease condition is already present and the differences between the case group with disease in comparison with a control group without the disease that is similar in all other ways possible (e.g., age, sex, comorbidities, community setting, etc.) are studied to help identify risk factors. In comparison with cohort studies, case control studies are relatively less expensive and shorter in duration and are mainly for explorative analysis. Cross-sectional studies focus on the analysis of data at one specific point in time. While case-control studies are concerned with a particular subpopulation, cross-sectional studies can include descriptive epidemiological data on an entire population at little expense. Case reports are detailed reports on individual cases including clinical presentation, diagnosis, treatment, and follow-up of the patient. Compared with the other studies described above, the case report is considered to be more anecdotal in nature and have significant limitations to objective evidence as it pertains to methodology and statistics. Nevertheless, case reports have useful roles by describing and documenting rare diseases or conditions, and unusual presentations of common diseases.

Diagnostic studies can be either RCTs or observational studies that are designed to establish the performance characteristics of a test or testing strategies. Alternatively, diagnostic studies may also aim at comparing two or more tests or testing strategies. To establish or compare performance, it is necessary to designate a reference method comparator as part of the study design. Performance of a new diagnostic test is ideally evaluated by comparison with a true reference method or "gold standard." In many cases, however, a true reference method does not exist and an imperfect standard must be used. Although statistical methods are well defined for assessing the accuracy of a new diagnostic test compared with a reference method, similar methods are sparse in a setting where a gold standard is unavailable. The challenge is to find a comparator as close as possible to the theoretical reference method, because accurate assessment of a new test may be compromised when an inappropriate reference strategy with its own uncertainty of estimations (e.g., the test is subjective, is histological, or contains findings of no clinical significance, etc.) is used. Further, the use of a reference standard may be limited if the new test is more advanced than the reference, or if the reference test is expensive, invasive, or unethical; while ensuring that the same reference standard is applied to all subjects to avoid bias.

Although diagnostic studies have traditionally focused on accuracy with a reference method, they do not address clinical outcomes. Indeed, RCTs would be ideal by randomizing patients and assessing whether or not the diagnostic test prior to treatment affects patient outcomes. Given that such studies are lacking or inadequate, analytical frameworks have been proposed and devised to include a "linked evidence approach" [9,10]. This analytical framework involves transferring or linking accuracy-based studies to RCTs on treatment effectiveness. Therefore a diagnostic test that is deemed accurate to a reference standard may be considered to be clinically effective if the population receiving the new test would be the same population that receives treatment for the condition. This transferability of the diagnostic study can be especially powerful if the RCT has shown good evidence for positive outcomes. Table 16.3 summarizes key domains that should be covered by SRs, RCTs, observational studies, and diagnostic test studies as well as their strengths and weaknesses.

Although EBLM has now evolved and placed more emphasis on SRs, narrative reviews still have a place in helping practitioners interpret evidence-based information as well as in instances where resources are lacking to perform a proper SR. Narrative reviews are usually written by experts in the field who have a keen interest in the review topic. This keen interest is frequently accompanied by a particular view (or bias) developed over years of study and often through direct involvement in the reviewed topic. On the one hand, narrative reviews can provide detailed insight into a field by an expert investigator and practitioner. On the other hand, narrative reviews are often developed using idiosyncratic methods and data presentation. For this reason, these reviews usually represent a subjective appraisal and qualitative summary of data that support the view of the expert author. Conflicting data are frequently discounted or not included at all. As a result, such reviews typically present a broad "10,000-foot" view about etiology, pathogenesis, diagnosis, clinical management, and the prognosis rather than addressing specific and focused clinical questions. Recognizing that narrative reviews are almost exclusively produced by proponents for test use is important. As such, narrative reviews tend to be subjective and are prone to bias and error, but are still useful for gaining background knowledge about a subject and may help with the dissemination of EBLM.

Indeed, utilization of clinical practice guidelines (CPGs) depends on the four As: physicians must be aware of them, agree with them, adopt them, and adhere to them [11]. Studies have shown that production and dissemination of evidence is not enough to ensure CPGs are put into practice. It has been suggested that there are leaks or barriers in the pipeline encompassing the four As, thus leading to diminished adherence and that proper investigation into each stage of CPG utilization is needed [12]. For proper success of guideline implementation, it is important to assess awareness, agreement, adoption, and adherence by clinicians through questionnaires and actual clinical practice audits, while further studying the causes of diminished adherence or increased barriers.

Critical appraisal of evidence

Critical appraisal is the systematic evaluation of clinical research papers to determine: (1) if the study addresses the question of interest in a focused way; (2) if valid methods were used to address the question; (3) if valid, are the study's results important to answering the question?; and (4) are the results applicable to the relevant patient or population? Fortunately, there are tools available for use in the Critical Appraisal process; a good example of these tools can be accessed at the Centre for Evidence-Based Medicine portal at http://www.cebm.net/critical-appraisal/. Here "Worksheets" in a checklist format have been developed that are specific for evaluating the type and design of study being assessed, that is, separate tools for diagnostic studies, SRs, prognostic studies, and therapeutic/RCTs. The worksheets guide the appraiser through the process with a series of steps. For example, for studies dealing with diagnostic questions, Step 1 assesses a series of items that address: are the results of the study valid?; Step 2 has items regarding: What were the results?; and Step 3 evaluates the applicability of the results (full listing of the diagnosis study tool is available at http://www.cebm.net/wp-content/uploads/2014/04/diagnostic-study-appraisal-worksheet.pdf).

Critical appraisal is frequently accomplished by a single individual whose goal is evaluation of evidence, addressing a question relevant to a local case or issue. However, when conducting SRs, typically two or more appraisers independently examine each source of evidence using a common assessment tool and scoring system. Table 16.6 displays such a scoring system that is used for assessment of evidence in the CDC's Laboratory Medicine Best Practice program [2]. Training of the appraisers is necessary to assure harmony in the criteria and reviewing methodology. Moreover, criteria must be established that defines when appraisers' scores are "discrepant" and explicitly states how discrepancies are resolved.

Critical assessment is also a task undertaken when deciding if a developed guideline should be championed or adopted/adapted by an institution, group, or professional organization. While a number of guideline appraisal tools are available, the second edition of the Appraisal of Guidelines for Research and Evaluation (AGREE) instrument [13] has been widely utilized by many clinical

TABLE 16.6 Summary table format for rating the body of evidence table.

Citation from bibliography	Study quality rating					
	Study[a] characteristics	Practice description	Outcome measures	Results/ finding	Total[a] quality points	Quality[b] effect size[c] rating (qualitative) (qualitative)
Study 1	0–3	0–2	0–2	0–3	Maximum 10	Good, Fair, or Poor,[b] Substantial, Moderate, etc.[c]
Study 2	0–3	0–2	0–2	0–3	Maximum 10	Good, Fair, or Poor,[b] Substantial, Moderate, etc.[c]
…	….	….	….	….	….	….

[a]*Range of values for each domain and total Quality Points from sum of the four domains; maximum 10-point scale.*
[b]*Conversion of total quality points to quality rating:* G, *Good, 8–10 total quality points;* F, *Fair, 5–7 total quality points;* P, *Poor, ≤ 4 total quality points.*
[c]*Effect size for each reference and study rated as: Substantial, Moderate, None/Minimal, or Adverse.*
Source: Adapted from R.H. Christenson, S.R. Snyder, C.S. Shaw, J.H. Derzon, R.S. Black, D. Mass, et al., Laboratory medicine best practices: systematic evidence review and evaluation methods for quality improvement. Clin. Chem. 57 (2011) 816–825.

and laboratory societies. The AGREE system addresses 23 key points that should be included in guidelines and the quality score of each. An analysis of using the AGREE II instrument has recently been published [14]. Whether the scope of a clinical question is on the individual, institutional, national, or international level, thoughtful and methodical critical assessment of the body of evidence is essential for ensuring the validity, significance, and relevance in the EBLM process.

Another tool intended to guide investigators in reporting studies of the diagnostic accuracy of laboratory tests and critically evaluate studies to detect bias is a checklist developed from the Standards for Reporting of Diagnostic Accuracy (STARD) initiative and STARD statement [14–16]. In this initiative, several editors of leading scientific and clinical journals reached a consensus statement on what elements regarding the performance of a laboratory test should be reported in studies of diagnostic accuracy. Since publication of the initial STARD statement, this initiative has led to significantly greater standardization in how laboratory test diagnostic accuracy studies are reported. An updated STARD checklist of essential items for reporting diagnostic accuracy studies was recently published [15]. This 2015 STARD checklist includes nearly 35 elements for reporting studies on the diagnostic accuracy of laboratory tests in the sections of title, abstract, introduction, methods, results, and discussion and for other information (Table 16.7). It is recommended that all studies on diagnostic accuracy should use the STARD checklist. One noteworthy point with the 2015 STARD document is that it mainly focuses on the validity of analysis. Appraisers must also evaluate the preanalytical and postanalytical aspects of the study, because these phases are most susceptible to error and are not included fully in the STARD checklist. In addition, readers may use the STARD checklist to assess the potential for bias and validity of the study conclusions.

Using the evidence to develop clinical practice guidelines

Clinical practice guidelines

Development of CPGs is a critical activity in using evidence to affect the practice of laboratory medicine and enhance the quality of patient care. CPGs provide statements that inform healthcare providers on how best to use evidence to make appropriate clinical decisions in specific patient care situations with a goal of optimizing the impact of guideline recommendations on patient outcomes in a way that is intended to improve patient care. Additional important goals for developing and using CPGs include the aim to disseminate best practice based on scientific evidence; decrease practice variation and the potential or frequency of professional misconduct; improve patient safety; improve the quality and effectiveness of care; improve cost-effectiveness of care; facilitate training, education, and continuous professional development; and increase explicitness, transparency, patient information, and autonomy of choice [17]. In laboratory medicine, guidelines also aim to improve the appropriateness of test utilization [18], an issue of growing significance for healthcare providers including clinical laboratory professionals.

While the developers of CPGs generally intend to achieve all or most of the above goals, experience has proven that there is a significant amount of variability from one CPG development group to another and even among a single CPG development group over time working on different CPGs. Fortunately, as more issues, variations and inconsistencies have been identified, numerous resources and tools have been created, developed, and refined toward standardizing CPG development processes and ensuring a higher likelihood that CPGs are developed in accordance with the best practices in guideline development. In recent years, these issues have been addressed

TABLE 16.7 Standards for reporting diagnostic accuracy studies: questions that should be addressed by a good diagnostic study.

Section and topic	#	Item	On page #
Title or abstract			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	
Abstract			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	
Introduction			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	
	4	Study objectives and hypotheses	
Methods			
Study design	5	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	
Participants	6	Eligibility criteria	
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, and inclusion in registry)	
	8	Where and when potentially eligible participants were identified (setting, location and dates)	
	9	Whether participants formed a consecutive, random, or convenience series	
Test methods	10a	Index test, in sufficient detail to allow replication	
	10b	Reference standard, in sufficient detail to allow replication	
	11	Rationale for choosing the reference standard (if alternatives exist)	
	12a	Definition of and rationale for test positivity cutoffs or result categories of the index test, distinguishing prespecified from exploratory	
	12b	Definition of and rationale for test positivity cutoffs or result categories of the reference standard, distinguishing prespecified from exploratory	
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	
Analysis	14	Methods for estimating or comparing measures of diagnostic accuracy	
	15	How indeterminate index test or reference standard results were handled	
	16	How missing data on the index test and reference standard were handled	
	17	Any analyses of variability in diagnostic accuracy, distinguishing prespecified from exploratory	
	18	Intended sample size and how it was determined	
Results			
Participants	19	Flow of participants, using a diagram	
	20	Baseline demographic and clinical characteristics of participants	
	21a	Distribution of severity of disease in those with the target condition	
	21b	Distribution of alternative diagnoses in those without the target condition	
	22	Time interval and any clinical interventions between index test and reference standard	
Test results	23	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	
	25	Any adverse events from performing the index test or the reference standard	
Discussion			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalizability	
	27	Implications for practice, including the intended use and clinical role of the index test	
Other information			
	28	Registration number and name of registry	
	29	Where the full study protocol can be accessed	
	30	Sources of funding and other support; role of funders	

STARD, Standards for reporting of diagnostic accuracy.
Source: Reproduced with permission from Clinical Chemistry; P.M. Bossuyt, J.B. Reitsma, D.E. Bruns, C.A. Gatsonis, P.P. Glasziou, L. Irwig, et al., STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clin. Chem. 61 (2015) 1446–1452.

as a top priority in the United States. Through the 2008 *Medicare Improvements for Patients and Providers Act*, Congress requested that steps be taken to implement recommendations from an IOM report on *Knowing What works in Health Care*. This, in turn, led to Congress commissioning the Secretary of the Health and Human Services to authorize the IOM to develop two evidence-based methodological standards—one on the development of CPGs and the other on standards for conducting SRs.

Best practices in guideline development

The IOM published *Clinical Practice Guidelines We Can Trust* in 2011 [19]. This report may be the single most comprehensive and authoritative resource presently available to those who intend to develop CPGs using the best practices that are presently recognized. Despite the variations in guideline development, which have long been recognized, particularly as numerous clinical organizations have developed their own procedures and processes for guideline development, the IOM report is built on a framework of standards for developing trustworthy CPGs. The eight critically important standards are listed in Table 16.8.

It may not always be possible to address every IOM standard in the development of a CPG. But the body of evidence regarding best practices in guideline development underscores that CPG developers should work toward significant adherence to these standards to the greatest extent possible. In fact, there is presently an abundance of resources and information available to developers of CPGs. Many professional organizations and societies have established guideline development policies and processes based, in large part, on the IOM report and standards with the end product being a standard operating procedure (SOP) for developing guidelines. Another group providing a wealth of resource information on guideline development is the Guidelines-International-Network (G-I-N). The G-I-N library contains many resources and materials that can be accessed at http://www.g-i-n.net/library.

Steps in guideline development

The first step in the development of any CPG is to select the topic of focus as well as how the subject area will be identified and refined. CPGs in the field of laboratory medicine can address diseases or conditions, symptoms, signs, or procedures (either therapeutic or diagnostic interventions), including the utility of tests for screening and monitoring. After identification of the exact area to be addressed, refining the scope of the guidelines should

TABLE 16.8 Institute of Medicine standards for developing trustworthy clinical practice guidelines.[a,b]

IOM Standard	Explanation
1. Establishing transparency	The process for guideline development including any sources of funding must be transparent, public, and fully disclosed
2. Management of COI	COI must be declared prior to CPG developing groups being formed. Leaders of these groups should not have any real or perceived COI. Sources of CPG funding must not influence the content of CPGs
3. Guideline development group composition	CPG group composition should be multidisciplinary including all stakeholders as well as patients and, if applicable, consumers
4. CPG—systematic review intersection	To the greatest extent possible, CPGs should be based on systematic reviews s that have been developed in accordance with recognized best practices
5. Establishing evidence foundations for and rating strength of recommendations	Recommendations should have clear descriptions of potential harms and benefits with a summary of evidence for each. Strength of evidence and strength of recommendations must be graded with differences of opinion explicitly stated
6. Articulation of recommendations	Recommendations must be clear and unambiguous in such a way as to allow full evaluation of compliance
7. External review	Draft CPGs should be provided to all relevant key stakeholders including the public for external review. CPG developers should record and address all comments regarding whether the issues raised were incorporated or not in the final CPG recommendations
8. Updating	Dates for the systematic review, CPG final publication, and future updates of the CPG should be documented in the initial final CPG. CPG developers should have a process identified for regular review of new evidence that could modify or alter future recommendations

COI, Conflict of interest; *CPG*, clinical practice guidelines; *IOM*, Institute of Medicine.
[a]*Explanation of the eight IOM standards adapted, in part, from reference [16].*
[b]*Standards for developing trustworthy CPGs from reference [18].*

follow. What are the critical issues to be addressed and which ones can be omitted? Which issues are important clinically? Where do laboratory professionals or clinicians using the laboratory need guidance? Does the group developing the guidelines have enough expertise or should others be brought on board?

The initial step in the formation of the guideline development group is to select a chair and usually a co-chair or vice-chair. The composition of the group overall should include experts on the procedures and conditions from areas of laboratory medicine and clinical practice relevant to the focus of the CPG subject matter. Additional experts may be included to bring supplementary expertise in special areas such as statistics, appraising of evidence, and implementation of guidelines. Staff members from involved clinical organizations may be available to the group to assist in retrieving evidence, organizing meetings of the group, and coordinating publication of guidelines. When more than one organization or society collaborates on the development of a CPG, it is imperative that every organization involved have at least one member of the guideline development group. As noted in the IOM standards, a representative from the public or, if applicable, the consumer base is also desirable. All members of the guideline development group should fully disclose any potential conflict of interest (COI) in the earliest stages of the group's efforts. Chairs and vice-chairs of groups developing CPGs must not have any COI.

CPGs should be based on an SR of the literature with the provision of criteria for identifying the quality of the evidence and the strength of each recommendation. An established, clear, explicit, and transparent system for grading the recommendations and weighting the evidence must be preestablished. Recommendations established by consensus are acceptable as long as this approach is explicitly described and those recommendations of relevance clearly identified. However, before any SR is initiated, the CPG development group must define the CPG topic and focus including potential approval by an oversight leadership group in one or more professional societies.

Fundamentals for conducting a solid, evidence-based SR are described in several resources including the IOM report on conducting SRs [6]. The CPG group must identify the background for the search, the key question(s) based on PICOTS criteria, and the analytic framework for how the literature search process will address the PICOTS elements including the specification of outcomes to be evaluated. Other issues the CPG development group will likely need to address include defined conditions and target populations and identification of the primary target audience for which the CPG is intended. Many other issues must be addressed. But some of the most significant issues to consider are the impact of the evidence report on technology assessment and associated costs to the healthcare system, how the impact of the CPG recommendations will be monitored, and, typically, what the budget and timelines are for CPG development.

It is recommended that guidelines be reviewed by experts in the clinical content area, experts on systematic reviewing and guideline development, and potential users of the guidelines. Because guidelines become outdated in time, determining a date when the guidelines will be reviewed and updated is important. Alternatively, "living guidelines" where systematically developed and evidence-based recommendations are updated as evidence becomes available can be considered. The establishment of guidelines also provides a thorough review of data on the utility of a test, the way in which the test may be used clinically, and the benefit derived from its use. The holistic nature of a guideline also helps to identify the way in which practice may need to change when implementing a novel test or intervention. A guideline also provides an overview of the quality of the evidence, which indicates the robustness of the data on effectiveness of a test and how wide the applicability of the test might be in different clinical settings. Much of the material for a guideline will have been generated from the SR of the peer-reviewed literature. However, guidelines must allow for potential applications that are unsupported by evidence, because primary studies are not always available to answer every clinical question

Additional resources for developing clinical practice guidelines

The 2011 IOM report is presently one of the most comprehensive resources available to those developing CPGs. The largest source of laboratory medicine practice guidelines (LMPGs) is also available on the website of the Academy of the American Association for Clinical Chemistry (AACC), formerly named the National Academy of Clinical Biochemistry (http://www.nacb.org). While most of the Academy's earlier efforts are now archived, guidelines are available at this website such as those focusing on the use of tumor markers in selected cancers, and laboratory analysis and application of pharmacogenetics in clinical practice, pain management, biomarkers of cardiac disease, and diagnosis and management of diabetes mellitus.

Another resource found on the AACC's Academy website is a recently revised SOP for preparing, publishing, and revising LMPGs or CPGs. It also includes procedures for review and approval of external society/organization guidelines for endorsement and support by AACC and its Academy. Additional selected useful

resources for developing CPGs including the AACC Academy SOP are listed in Table 16.9. While this list is not intended to identify all resources available, tools such as GRADE, Guideline Implementability Appraisal, AACC's Academy SOP, STARD, Preferred Reporting Items for Systematic Reviews and Meta-Analyses, and AGREE II are well-recognized tools in the field of CPG development, particularly for guidelines relating in some manner to the use of clinical laboratory tests. The strengths and limitations of AGREE II are specifically discussed in the 2011 IOM report. Several CPG developers, including the AACC Academy, presently use the AGREE II resource to improve their guidelines while in development with a goal of striving for best practices.

Implementation of clinical practice guidelines

As stated previously, there are leaks or barriers in the CPG pipeline and for proper success of guideline implementation, and it is important to assess awareness, agreement, adoption, and adherence by healthcare practitioners. Indeed, production of good quality CPGs does not guarantee the implementation of best practices. Thus more focus must be put on guideline use instead of solely on guideline development. Guideline adaptation can be done in four possible ways: (1) adopting the recommendations without modification; (2) adapt recommendations to a new context; (3) develop recommendations de novo based on existing reviews of evidence; or (4) develop recommendations de novo based on new evidence [21]. It is often that the recommendations from guidelines that pertain to public health are more difficult to implement since this includes the use of entire health-care systems and many stakeholders [21]. Thus it is important to account for the local context for successful guideline implementation.

Given the barriers to implementation, current movements now recognize that guideline adaptation is more defined as a systematic approach to considering the use and/or modification of CPGs produced in one cultural and organizational setting to be applied in another different context [22,23]. In other words, adaptation is an extension of guideline development, whereby the existing CPG is customized to ensure the recommendations are suited to the local needs, priorities, legislation, policies, and available resources. Adaptation of guidelines can occur either formally or informally. Informal adaptation occurs without an established framework and may include at an institutional level where guidelines are adapted according to the AGREE instrument [21]. Informal adaptation can also occur on an individual basis where the healthcare provider adapts practice guidelines according to individual patient needs and resources such as the availability of certain laboratory tests or therapy [21]. The disadvantage of informal adaptation is that it can veer significantly outside the scope of the original recommendation. On the other hand, informal adaptation is performed according to an established framework to address the complexity, increase the rigor of review, and enhance the quality of the adapted guideline. To address these needs, the ADAPTE collaboration was born of two independent groups focusing on guideline adaptation: the ADAPTE group and the Practice Guideline Evaluation and Adaptation Cycle. The ADAPTE collaboration convened as an international collaboration of researchers, guideline developers, and guideline implementers to promote the development and use of CPGs through the adaptation of existing guidelines.

The main objective of this group was to develop and validate a generic user-friendly adaptation process that maintains the validity and high quality of the adapted CPGs as well as promote the users' sense of ownership of the adapted CPG. The ADAPTE manual includes a stepwise approach to guideline adaptation (Table 16.10), and there is an active focus on the barriers to guideline implementation and facilitating factors. Although the ADAPTE method employs a more formative, instead of prescriptive, approach, it should be realized that guideline implementation is a long process taking more than years in some cases and more complex than anticipated, particularly in the absence of guideline methodology and facilitation

TABLE 16.9 Additional resources for developing clinical practice guidelines.

Tool and link	Purpose
AACC Academy SOP (http://www.nacb.org)	Standard operating procedures for developing LMPGs
STARD 2015 (Ref. [15])	Updated checklist for studies on the diagnostic accuracy of laboratory tests
GRADE (http://www.gradeworkinggroup.org)	Grading the strength of evidence and recommendations
GLIA (http://nutmeg.med.yale.edu/glia)	Useful procedures and information for implementation of guidelines
AGREE II (www.agreetrust.org/) (also Refs. [13,14,20])	Useful tool for critical appraising guidelines in development and in their final stage addressing several IOM standards

AACC, American Association for Clinical Chemistry; *AGREE*, Appraisal of Guidelines for Research and Evaluation; *IOM*, Institute of Medicine; *GLIA*, GuideLine Implementability Appraisal; *LMPG*, Laboratory Medicine Practice Guidelines; *SOP*, standard operating procedures; *STARD*, standards for reporting of diagnostic accuracy.

TABLE 16.10 The ADAPTE process.

Phases	Tasks	Associated modules
Set-up phase	Prepare for ADAPTE process	Preparation
Adaptation phase	Define health questions	Scope and purpose
	Search and screen guidelines	Search and screen
	Assess guidelines	Assessment
	Decide and select	Decision and selection
	Draft guideline report	Customization
Finalization phase	External review	External review
	Plan for future review and update	Aftercare planning
	Produce final guideline	Final production

Source: Adapted from Collaboration A. Guideline adaptation: a resource toolkit, version 2.0, 2009.

support [21,24]. This is particularly important for groups that are newly formed and that are inexperienced in undertaking guideline adaptation and implementation planning. Nevertheless, a major benefit of guideline implementation is that this process engages front-line care workers and allows them to take ownership of the best practices and recommendations outlined in the CPGs [24]. Given the difficulties of guideline implementation, the CAN-IMPLEMENT framework has recommended that this process should address: (1) engagement and capacity building; (2) adaptation of a high-quality CPG; (3) the continuum of evidence-informed practice (as opposed to an episodic activity); and (4) a planned-action framework as a roadmap to embed properly and effectively guideline adaptation [24].

Implementation of evidence-based guidelines can lead to better patient outcomes by helping healthcare professionals to adopt best practices and assist policymakers in designing effective programs. It is estimated that 30%–40% of patients receive treatment that is not based on scientific evidence and that 20%–25% receive treatments that are not needed or potentially harmful [25]. Therefore the proper implementation of CPGs must be considered or else risk patient harm. Furthermore, proper implementation of CPGs according to established frameworks can prevent the waste of financial and human resources that are consumed in the development of guidelines. However, there remain several gaps in knowledge about guideline implementation. It is relatively unknown how adaptation of guidelines differs between countries with differences in economic development, since low-income countries may have limited resources. Formal frameworks have not been adequately evaluated in terms of their quality and have been mostly subject to subjective self-administered evaluations. Furthermore, formal evaluations have not been done to determine whether methodological experts could improve the adaptation framework. It is also unclear whether the expedited approach to adapt guidelines is worth the expense and whether this can affect the validity of the resultant recommendations of the original guideline. Finally, formal frameworks are in need of a process that advises guideline developers on how to implement the guideline.

Metrics for evaluating diagnostic laboratory tests

Measures of diagnostic accuracy

(Some concepts included in this section are also covered in Chapter 2: Statistical methods in laboratory medicine, and Chapter 4: Method validation, and the reader is directed to these sections for further reference.)

Diagnostic specificity and sensitivity are two common statistical indices that have been defined to aid in the evaluation of qualitative or quantitative tests. Sensitivity of a test is defined as its ability to make a positive diagnosis in patients who have the target disease. Specificity of the test is defined as its capability to make a negative diagnosis in patients who do not have the target disease. Both sensitivity and specificity are commonly expressed as percentages or decimals. As illustrated in Table 16.11, the sensitivity of the test is the number of true-positive results divided by the total number of cases with a confirmed presence of the disease. Specificity of a test is the number of true-negative cases divided by the total number of cases with confirmed absence of the disease. Frequently, sensitivity, and specificity are considered to be independent of the disease prevalence within a population. In reality, however, this supposed independence is not strictly true, because the definition of a clinical question often specifies the presence of clinical variables such as signs and symptoms. In even a slightly different population, for example, an urban versus a rural setting, sensitivity, and specificity may vary, because the proportion of false negatives and/or false positives may be different.

TABLE 16.11 2 × 2 contingency table for evaluating accuracy of diagnostic tests.

Result	Disease present	Disease absent	Total
Test positive	TP	FP	TP + FP
Test negative	FN	TN	FN + TN
Total	TP + FN	FP + TN	TP + FN + FP + TN

Notes: Sensitivity = TP/(TP + FN); specificity = TN/(TN + FP); positive predictive value = TP/(TP + FP); negative predictive value = TN/(TN + FN). *FP*, False-negative; *TP*, true-positive.

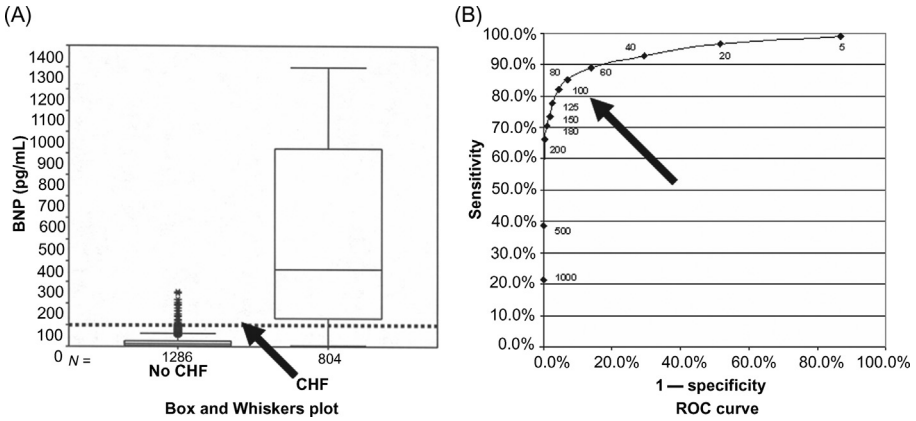


FIGURE 16.3 (A) Box and whiskers plot showing patients with target condition (congestive heart failure) and no disease (no congestive heart failure). The top of the box is the 75th percentile of data, and the bottom of the box is the 25th percentile. The line in the box is the median, and the error bars represent the range of data. (B) Receiver operating characteristic curve. Receiver operating characteristic curves are used to show the pattern of sensitivities and specificities when the performance of the test is evaluated at several diagnostic thresholds. The overall diagnostic performance of a test can be evaluated by examining the area under the curve. The *arrow* indicates the decision point that corresponds to the arrow at the cutoff point line in Panel A. This decision point yields the least number of false positives and false negatives.

Positive predictive value (PPV) and negative predictive value (NPVs) are also important performance characteristics. PPV expresses the probability that a patient has a target disease given that the test's result is positive [TP/(TP + FP)]. NPV expresses the probability that a patient does not have the target disease given that the test's result is negative [TN/(TN + FN)]. (See a more detailed discussion calculation in Chapter 2: Statistical methods in laboratory medicine.) PPV and NPV are parameters that are usually of interest to clinicians, because the following questions are frequently relevant: "What is the probability that the patient has the target condition when the test is positive (PPV)?" and/or "How often does a negative test result rule out the target condition (NPV)?" PPV and NPV depend on the prevalence of disease in the population examined. Although sensitivity and specificity of the test for the target condition may be high, the performance of the test for answering the clinician's question about the meaning of a positive test result may be poor if the disease prevalence is low. This is because a disproportionate number of positive results will be false-positives, and the positive result may not confer a high probability of the target condition. The dependence of predictive value on disease prevalence often confounds translation of clinical trial results and the test performance in clinical practice. Thus results of clinical trials should be considered the most optimistic performance of the test, because the disease prevalence and "all-comers" population specifics in "real-life" practice may not exactly match that in the trial.

Receiver operating characteristic curves

Receiver operating characteristic (ROC) curves are a plot of sensitivity on the *y*-axis versus (1 − specificity) on the *x*-axis, as shown in Fig. 16.3B. ROC curves always confer a binary outcome, for example, "yes" or "no" for a diagnosis, "yes" or "no" for an adverse event, "alive" or "dead" as an outcome, and so on. ROC curves reflect the performance of a test by evaluating the sensitivity and specificity of the binary outcome at a continuum of diagnostic thresholds or cutoffs from the actual plot of data (Fig. 16.3A). The overall performance of the test can be determined from the area under the ROC curve (Fig. 16.3B). Poor or useless tests have ROC curve areas close to 0.50, and the plot appears as a diagonal where sensitivity = 1 − specificity. For an excellent test, the

ROC curve rises steeply and passes close to the upper left-hand corner; the perfect test has an area of 1.0, where both the sensitivity and specificity are 100%. ROC curve areas should always be reported with a confidence interval around the area's value; commonly, the 95% confidence interval is utilized.

ROC curves are particularly useful for comparing the performance of two tests; this is done by running the tests on a single cohort of patients, and the test showing a significantly larger area is the better test. If there is no significant difference between the two, then diagnostic performances of the tests do not differ in their ability to predict the outcome. In SRs, summary ROC curves are used to display the results of a set of studies. In such plots, the sensitivity and specificity from each study are plotted as a separate point on the summary ROC curve.

Likelihood ratios

After formulating a clinical question—for example, "Does my patient have a target diagnosis?"—the clinician must estimate the pretest probability. Pretest probability and the pretest odds are both directly related to the prevalence of the target condition in patients showing similar signs and symptoms according to the following expression:

$$\text{Pretest probability} = \text{Prevalence of the target condition} \tag{16.1}$$

To convert into pretest odds,

$$\text{Pretest odds} = \frac{\text{Prevalence}}{1 - \text{Prevalence}} \tag{16.2}$$

Clinicians request laboratory medicine services after the pretest probability has been estimated, to assist in either raising or lowering the likelihood that the target condition is the diagnosis for their patients. This is done by determining the likelihood ratio, which is calculated to add insight into how likely it is that the patient has the target condition if the test is positive (positive-likelihood ratio) or how likely it is that the condition can be ruled out if the test is negative (negative-likelihood ratio). For practical purposes, the positive or negative likelihood ratios can be easily derived from diagnostic sensitivity and specificity according to the following.

$$\text{Positive} - \text{Likelihood ratio} = \frac{\text{Sensitivity}}{1 - \text{Specificity}} \tag{16.3}$$

$$\text{Negative} - \text{Likelihood ratio} = \frac{1 - \text{Sensitivity}}{\text{Specificity}} \tag{16.4}$$

The positive-likelihood ratio is interpreted as the likelihood that the patient has the target condition if the test is positive. The negative likelihood ratio is interpreted as the likelihood that the patient has the target condition if the test is negative.

The posttest odds of the patient having the target condition can be estimated once either a positive or negative test result is known, and the respective likelihood ratio calculated.

$$\text{Posttest odds} = \text{Pretest odds} \times (\text{Positive or negative}) \text{ likelihood ratio}$$

The posttest probability for the presence of the target condition if the test was positive or negative can be estimated from the following:

$$\text{Posttest probability} = \frac{\text{Posttest odds}}{1 + \text{Posttest odds}} \tag{16.6}$$

When positive, an excellent test will improve the pretest probability to a posttest probability value approaching 1.0 for target condition; when negative, the test will virtually rule out the condition. In practice, few laboratory medicine tests achieve excellent performance.

An excellent tool to facilitate determination of posttest probabilities is termed the Fagan nomogram [26] and is displayed in Fig. 16.4. This diagram allows simple conversion by simply connecting a line from the pretest probability (or prevalence) through the likelihood ratio (positive or negative, depending on the test result) to estimate the posttest probability.

Test as a probability modifier: example

A clinician asks "What is the probability that my patient with dyspnea in an emergency department (ED) setting has a diagnosis of congestive heart failure (CHF) if the B-type natriuretic peptide (BNP) test is negative?" Given that the prevalence in the ED population is 25%, the pretest odds = prevalence / (1 − prevalence) = 0.25/0.75 = 0.33. If the sensitivity of BNP for CHF in ED patients with dyspnea is 93% and the specificity is 85%, then the negative likelihood ratio = 0.07 / 0.85 = 0.08. Thus the posttest odds = pretest odds • negative likelihood ratio = 0.33 × 0.08 = 0.026. Converting posttest odds into posttest probability = posttest odds / (1 + posttest odds) = 0.026 / (1 + 0.026) = 0.025.

Thus only ∼2.5% of patients with dyspnea in the ED with a negative BNP test will have a diagnosis of CHF. Use the Fagan pretest/posttest probability nomogram in Fig. 16.4 to check this posttest estimation.

Is the test cost-effective?

There are three main approaches in health economic evaluations: cost−benefit analysis (CBA), cost-effectiveness analysis (CEA), and cost−utility analysis (CUA).

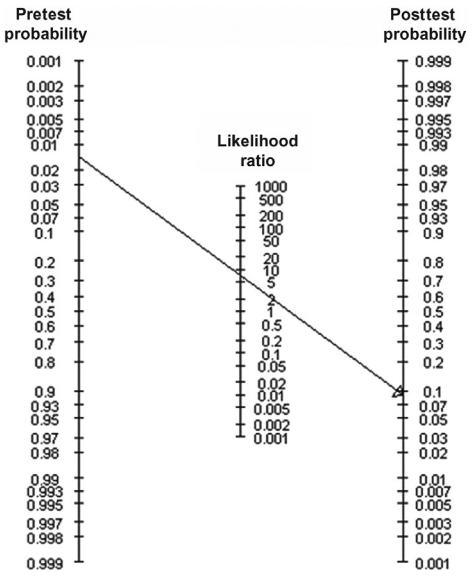


FIGURE 16.4 Fagan pretest/posttest probability nomogram [26]. To determine posttest probability, connect a line from the pretest probability value through the likelihood ratio for the diagnostic test and extending to the intersection with the right-hand axis to estimate the posttest probability. As an example, a line has been connected for a test with a pretest probability of 0.015 through a likelihood ratio of 7.5 to estimate the posttest probability of 0.1.

Although all three approaches require measurement of both cost and benefit, they differ in how effectiveness is measured and whether or not effectiveness is valued in terms of money.

Cost–benefit analysis

The goal of CBA is to determine whether the benefits of an intervention exceed its costs. CBA is suited to answering the question "Is the intervention worthwhile?" One challenge of CBA is that health outcomes must be valued by assigning a dollar amount to a year of life. Valuation of health benefits may be achieved by using the human capital approach, where the value of an intervention is based on the impact on a person's productivity and the value of earning based on wage rates. Alternatively, valuation of health benefit can be achieved through the willingness-to-pay (WTP) approach, in which respondents to surveys are asked to state the maximum amount they are willing to pay for an intervention, illustrated in the example below [27]:

> *Scenario*: Assume that your risk of cervical cancer death is the same as a woman living in Pennsylvania and screened regularly (1 in 37,000). By having the new Pap test, the risk of dying from cervical cancer is reduced to 1 in 50,000.
> *Question*: What is your WTP out-of-pocket for the new Pap smear test?
> *Choices*: \$0, \$200, \$300, \$400, \$500, \$600, \$700, \$800, \$900, \$1000.
> *Responses*: 175 females, average age 39 years, mean WTP \$237.00.

Cost-effectiveness analysis

CEA is aimed at answering the question "Given a fixed budget, what is the most efficient way (maximizing effectiveness) to spend it?" In CEA, effects are measured in one single metric that is a natural unit. Cost-effectiveness could be measured in the cases of disease averted or change in a physiologic measure. A preferred metric is the one that can be used to compare across diagnostic tests for different diseases. For example, life years or survival is a long-term, downstream measure of outcome that can be used to compare a new liquid-based cytology test for cervical cancer screening [28] and fecal occult blood testing for colorectal cancer screening [29]. The incremental cost per life year saved by using the new technology for cervical cancer screening was approximately \$147,000 with triennial screening [28]. Because the outcome measure metric is cost per life year saved, the cost-effectiveness of cervical cancer screening can be compared directly with the cost-effectiveness of colorectal cancer screening at \$92,000 per life year saved for annual fecal occult blood testing plus sigmoidoscopy [29].

Cost–utility analysis

In this approach, an attempt is made to consider the quality of the health outcome in addition to the quantity of the health outcome. Thus years of life in states less than full health are converted into healthy years based on health-state preference values. The result is usually expressed as quality-adjusted life years [30].

Economic evaluations provide the tools necessary to shift the way practitioners think about diagnostic tests and their value in the context of competing demands for healthcare dollars. An economic evaluation requires the identification, measurement, and valuation of the clinical benefits of a new intervention and the associated costs. There are well-accepted criteria for assessing economic evaluations that also apply to economic evaluations of diagnostic tests, and studies should be critically reviewed and planned using these criteria as a guideline [31].

Using clinical laboratory medicine data for decision-making

In the practice of laboratory medicine, a multitude of factors and considerations contribute to the value of laboratory test results. These well-known factors include the prevalence of disease in the population as well as the technical performance, diagnostic accuracy, and other considerations that largely focus on the laboratory test

itself. Furthermore, within the context of EBLM, this hierarchy of factors should always start with an appropriate clinical question that provides a focused framework for the assessed use of the laboratory test.

Consistent with the IOM's six aims for health care, a laboratory test result realizes its fullest potential and value when it truly affects patient-centered care, that is, there is a measurable impact on patient outcomes. Years ago, performing a laboratory test that was analytically robust with excellent diagnostic sensitivity and specificity would usually be sufficient for adequate reimbursement from most payers. But presently, while these factors are still requisite characteristics of a laboratory test, a growing awareness in providers and payers alike is that this same laboratory test result has considerably more value when it affects patient outcomes. As a result, the recently changing landscape of health care has defined a new paradigm for how the value of a laboratory test should be viewed. This paradigm can be illustrated in the laboratory test value-based pyramid depicted in Fig. 16.5.

As shown in Fig. 16.5, a laboratory test's technical performance, diagnostic accuracy, and appropriate clinical use remain fundamentally important. Furthermore, conducting various types of reviews of relevant laboratory medicine studies through appropriate practices is still required to produce significant EBLM recommendations. But these recommendations become even more significant when they can demonstrate or lead to showing how the laboratory test changes clinical practice and affects patient outcomes. The importance of patient outcomes was addressed in the Affordable Care Act through the creation of the Patient-Centered Outcomes Research Institute (PCORI). The purpose of PCORI was to help patients, clinicians, and policymakers make well-informed decisions about health care considering findings from comparative effectiveness research studies [32]. More information about PCORI can be found at www.pcori.org.

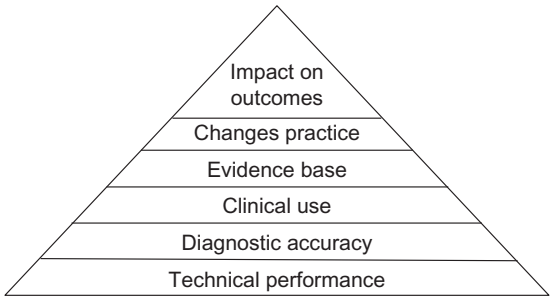


FIGURE 16.5 The laboratory test value-based pyramid. The value of a laboratory test still rests on the foundation of technical performance and diagnostic accuracy as well as appropriate clinical use. Ideally, these characteristics lead to the building of a robust evidence base. But in today's healthcare environment, the greatest value of a laboratory test is only fully realized if use of the test as supported by the evidence base changes practice and impacts patient outcomes.

But the critical importance of changing clinical practice and impacting patient outcomes sets a higher bar for even the most robust EBLM studies and activities. In the present healthcare environment, it is more important than ever to be able to translate the evidence into action. If successful in this effort, the final step is to understand and be able to use the presently available quality indicators and metrics that are applicable to the laboratory test value context being considered.

From evidence to action: changing clinical practice and improving outcomes

Translating evidence into action is highly valued although how best to address many challenges remains an area of discussion and further investigation. Evidence that affects clinical practice comes from health research that is adequately funded in order to translate knowledge into evidence-informed practices and policy [33]. For the effective practice of patient-centered laboratory medicine, key areas needing to be addressed include utilization of existing and new laboratory tests, wider development of standardized protocols for conducting sound, and robust biomarker studies as well as the benchmarking of existing and new tests in specified situations using commonly accepted measures of effectiveness [34]. As discussed earlier in this chapter, developing each new or revised LMPG by AACC's Academy requires systematic and standardized processes that remain current with the best practices in guideline development. How AACC's Academy and the EBLM Committee (EBLMC) use the SOP developed LMPGs and CPGs that can translate into improved evidence-based practice and patient care is illustrated in Fig. 16.6 [35].

As the paradigm for using resources such as laboratory testing shifts from volume to value, use of new tools and strategies for gathering, analyzing, and applying evidence has become prevalent. Evidence-based reviews focusing on good standards of laboratory practice using published literature as well as results of unpublished quality studies are the approach used in the CDC's LMBP initiative [1,2]. Published reports from the CDC's LMBP initiative also demonstrate how evidence from published as well as unpublished studies can be combined in a meta-analyses summary forest plot to illustrate when evidence-based recommendations might be worthy of strong consideration for translation into practice. Going back to Fig. 16.2, studies were rated as fair or good using the point system previously described in Table 16.6 earlier in this chapter, and go on to demonstrate that the random effects means of fair studies, good studies, and all studies; overall, favor the use of barcoding system to reduce patient specimen identification errors [7].

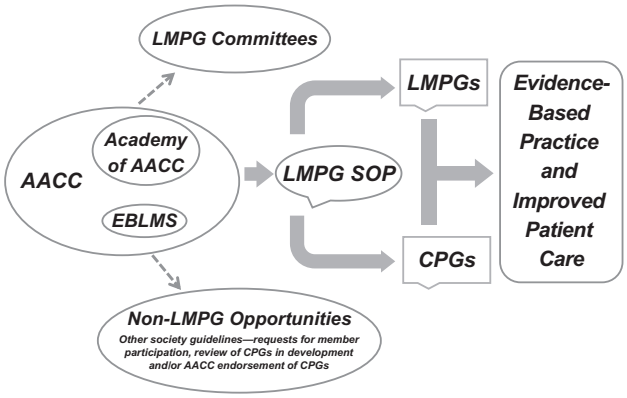


FIGURE 16.6 How the Academy and the Evidence-Based Laboratory Medicine Committee promotes evidence-based practice and improved patient care through use of Standard Operating Procedures for the development of laboratory medicine practice guidelines and clinical practice guidelines. *Source: Modified with permission for the e-Journal of the International Federation of Clinical Chemistry and Laboratory Medicine. The original article, Defining the path forward: guidance for laboratory medicine guidelines (http://www.ifcc.org/media/322288/eJIFCCVol26No3_pp158-167.pdf), was published in the e-Journal of the International Federation of Clinical Chemistry and Laboratory Medicine Volume 26 (3), p. 163, July 2015 [35].*

In addition to conducting SR or narrative review, other alternatives combine rigorous SR and meta-analyses with perspectives in a narrative format focusing on the best practices in the field by a consensus of experts. In a recent AACC Academy LMPG on laboratory analysis in the diagnosis and management of diabetes mellitus, roughly 25% of the final recommendations are "good practice points" (GPPs) [36]. GPPs are recommendations mostly driven by expert consensus and professional agreement and are based on specific items of information and/or professional experience, or widely accepted standards of best practice. In addition to [36], more information on the use of GPPs is available at https://www.aacc.org/science-and-research/practice-guidelines/diabetes-mellitus.

As laboratory professionals strive to apply the evidence from EBLM guidelines to their own practice settings, demonstrating value of laboratory test results for changing clinical practice and improving outcomes increasingly involves other medical disciplines and/or operating departments in many provider settings. A good example is the CER on the use of PCT to guide antibiotic therapy in selected patient populations [3]. This CER concluded that PCT-guided antibiotic therapy can lead to significant reductions in antibiotic use without adversely affecting patient outcomes in critically ill ICU patients. PCT-guided antibiotic therapy also reduced antibiotic prescription rates, and the duration of antibiotic therapy is selected patient groups such as those with respiratory tract infections without an increase in morbidity or mortality.

At present, one of the most intensively studied issues regarding the value of laboratory testing is that of laboratory test utilization. The accumulation and publication of evidence on this subject are accelerating, as numerous studies are being reported focusing on improving laboratory utilization. Excessive and inappropriate use of laboratory testing decreases its value. One recent meta-analysis conducted over a 15-year period reported that inappropriate testing takes many forms, may be as high as 30%–40% of all laboratory testing and involves overutilization as well as underutilization [37]. Taking effective action on recommendations from EBLM studies on laboratory utilization is a key priority for healthcare providers with many recognizing that "less is more for patients in laboratory testing" [38]. The American Board of Internal Medicine Foundation's website, www.choosingwisely.org, lists dozens of clinical society top recommendations to affect clinical practice and improve patient outcomes [39]. In the "Choosing Wisely" campaign, more than 40 (roughly 10%) of the evidence-based recommendations from all of the participating clinical societies pertain to improved use of laboratory tests. Table 16.12 lists a handful of key "Choosing Wisely" recommendations on laboratory tests and the societies that made them. As with all CPGs, laboratory professionals should apply the concepts of EBLM covered earlier in this chapter by using tools for critical appraisal and for adaptation and implementation of guidelines.

Quality indicators and metrics

It is said that quality improvement does not occur unless metrics associated with key aspects of the intervention are established and monitored on a regular basis. It is also said that quality improvement occurs more rapidly if monitoring is more frequent. In keeping with this notion, an important element of assuring that implemented practices and practice changes have the intended impact is establishing benchmark metrics for monitoring and auditing performance. Table 16.13 lists a number of organizations that have established quality indicators or metrics for improving healthcare venues and clinical care.

Summary

The processes and methods of EBLM are useful for determining what works for a continuum of healthcare stakeholders, from patients to practitioners to payers. The EBLM process consists of careful and proper question formulation, finding evidence, appraising the strength of the evidence, applying the knowledge to laboratory practice, and auditing the results. The use of medical evidence has significantly evolved from adoption of subjective narrative

TABLE 16.12 Selected recommendations from American Board of Internal Medicine's "Choosing Wisely" campaign related to the utilization of laboratory tests.[a]

Professional Society or Organization	Recommendation
American Association of Blood Banks	Do not perform serial blood counts on clinically stable patients.
American Association for the Study of Liver Diseases	Do not repeat hepatitis C viral load testing outside of antiviral therapy.
American College of Physicians	In patients with low pretest probability of venous thromboembolism, obtain a high-sensitive D-dimer measurement as the initial diagnostic test; do not obtain imaging studies as the initial diagnostic test.
American College of Preventive Medicine	Do not routinely perform PSA-based screening for prostate cancer.
AMDA—The Society for Post-Acute and Long-Term Care Medicine	Do not obtain a urine culture unless there are clear signs and symptoms that localize to the urinary tract.
American Society for Clinical Pathology	Do not perform population-based screening for 25—OH Vitamin D deficiency.
American Society for Clinical Pathology	Do not test for myoglobin or creatine kinase-MB in the diagnosis of acute myocardial infarction. Instead, use troponin I or T.
Critical Care Societies Collaborative—Critical Care	Do not order diagnostic tests at regular intervals (such as every day), but rather in response to specific clinical questions.
Society of Hospital Medicine—Adult Hospital Medicine	Do not perform repetitive complete blood count and chemistry testing in the face of clinical and lab stability.

ABIM, American Board of Internal Medicine.
[a]*Selected recommendations from the American Board of Internal Medicine Foundations' "Choosing Wisely" campaign. Accessed online at www.choosingwisely.org.*

TABLE 16.13 Organizations that have established quality indicators and/or have set metrics in health care.

Organization establishing metrics	Abbreviation	Website
Centers for Medicare and Medicaid	CMS	https://www.cms.gov/
College of American Pathologists	CAP	http://www.cap.org/web/home
The Joint Commission	TJC	http://www.jointcommission.org/
American Heart Association	AHA	http://www.heart.org/HEARTORG/
Society for Cardiovascular Patient Care	SCPC	http://www.scpc.org/
Commission on Office Laboratory Accreditation	COLA	http://www.cola.org/
American Association of Bioanalysts	AAB	http://www.aab.org/aab/Certification-Qualifications.asp
Centers for Disease Control and Prevention	CDC	http://www.cdc.gov/
National Academy of Clinical Biochemistry	NACB	https://www.aacc.org/community/national-academy-of-clinical-biochemistry
American College of Cardiology	ACC	http://www.acc.org/#sort = %40fstartz32xdatetime86069%20ascending
American Association of Chest Physicians	AACP	https://www.chestnet.org/Guidelines-and-Resources
Most State Health Departments	------	

opinion to objective SRs. This continuous growth of scientific literature has become more challenging and a number of helpful tools have been developed to evaluate the evidence. Further challenges to the delivery of high quality of laboratory services include increasing workload, continuous stream of new tests and new technology, uncontrollable demand, claims of excessive unnecessary testing, and slow transition of evidence to practice. In particular, more attention needs to be addressed to the impact of laboratory test utilization on health outcomes [40], while trying to cope with the problems of continuing education and the difficulties associated with changing practice [41]. All of these issues are compounded by the poor quality of evidence on the utility of new tests, the inability to demonstrate improved health outcomes, and the delays in bringing data into the peer-reviewed literature. EBLM should be used as the platform for decision-making for management of individual patients,

establishing guidelines, and guiding policy. In order to ensure that the best evidence is used, collaboration with clinical colleagues is essential for understanding how the laboratory information will be used to optimize care delivered to patients and improving health outcomes.

References

[1] R.H. Christenson, Preamble: 'evidence in action: the laboratory medicine best practice initiative', Clin. Biochem. 45 (2012) 977–978.

[2] R.H. Christenson, S.R. Snyder, C.S. Shaw, J.H. Derzon, R.S. Black, D. Mass, et al., Laboratory medicine best practices: systematic evidence review and evaluation methods for quality improvement, Clin. Chem. 57 (2011) 816–825.

[3] Procalcitonin-guided antibiotic therapy. Comparative effectiveness review. No. 78, Agency for Healthcare Research and Quality (AHRQ), Rockville, MD, 2012.

[4] (AHRQ) AfHRaQ. Systems to rate the strength of scientific evidence. In: Services USDOHaH, Agency for Healthcare Research and Quality (AHRQ), Rockville, MD, 2002.

[5] G.H. Guyatt, D.L. Sackett, D.J. Cook, Users' guides to the medical literature. II. How to use an article about therapy or prevention B. What were the results and will they help me in caring for my patients? Evidence-based medicine working group, JAMA 271 (1994) 59–63.

[6] Morton S., Berg A., Levit L., Eden J. (Eds.) Finding What Works in Healthcare - Standards for Systematic Reviews. National Academies Press, Washington, DC, 2011.

[7] S.R. Snyder, A.M. Favoretto, J.H. Derzon, R.H. Christenson, S.E. Kahn, C.S. Shaw, et al., Effectiveness of barcoding for reducing patient specimen and laboratory testing identification errors: a laboratory medicine best practices systematic review and meta-analysis, Clin. Biochem. 45 (2012) 988–998.

[8] Higgins J.P., Green S. (Eds.) Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration, 2011.

[9] T. Merlin, S. Lehman, J.E. Hiller, P. Ryan, The "linked evidence approach" to assess medical tests: a critical analysis, Int. J. Technol. Assess. Health Care 29 (2013) 343–350.

[10] D. Samson, K.M. Schoelles, Chapter 2: medical tests guidance (2) developing the topic and structuring systematic reviews of medical tests: utility of picots, analytic frameworks, decision trees, and other frameworks, J. Gen. Intern. Med. 27 (Suppl. 1) (2012) S11–S19.

[11] D.E. Pathman, T.R. Konrad, G.L. Freed, V.A. Freeman, G.G. Koch, The awareness-to-adherence model of the steps to clinical guideline compliance. The case of pediatric vaccine recommendations, Med. Care 34 (1996) 873–889.

[12] S. Mickan, A. Burls, P. Glasziou, Patterns of 'leakage' in the utilisation of clinical guidelines: a systematic review, Postgrad. Med. J. 87 (2011) 670–679.

[13] The agree enterprise website - appraisal of guidelines for research and evaluation (agree) instrument. <www.agreetrust.org>.

[14] Brouwers M.C., Kho M.E., Browman G.P., Burgers J.S., Cluzeau F., Feder G., et al. Agree II: advancing guideline development, reporting and evaluation in health care. CMAJ 2010;182:E839-E842.

[15] P.M. Bossuyt, J.B. Reitsma, D.E. Bruns, C.A. Gatsonis, P.P. Glasziou, L. Irwig, et al., Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies, Clin. Chem. 61 (2015) 1446–1452.

[16] P.M. Bossuyt, J.B. Reitsma, D.E. Bruns, C.A. Gatsonis, P.P. Glasziou, L.M. Irwig, et al., Towards complete and accurate reporting of studies of diagnostic accuracy: the stard initiative. Standards for reporting of diagnostic accuracy, Clin. Chem. 49 (2003) 1–6.

[17] A. Horvath, Are guidelines guiding us on how to utilize laboratory tests? eJIFCC 26 (2015) 146–157.

[18] A.R. Horvath, From evidence to best practice in laboratory medicine, Clin. Biochem. Rev. 34 (2013) 47–60.

[19] Institute of Medicine (US), Clinical Practice Guidelines We Can Trust. Washington, DC: National Academies Press, 2011.

[20] S.E. Kahn, J.R. Astles, S.F. Lo, M.J. Bennett, The AGREE II instrument is helpful for creation of national academy of clinical biochemistry laboratory medicine practice guidelines, Clin. Chem. 59 (2013) 446–447.

[21] Z. Wang, S.L. Norris, L. Bero, The advantages and limitations of guideline adaptation frameworks, Implement. Sci. 13 (2018) 72.

[22] Collaboration A. Adapter resource toolkit for guideline adaptation, version 1.0, 2007.

[23] Collaboration A. Guideline adaptation: a resource toolkit, version 2.0, 2009.

[24] M.B. Harrison, I.D. Graham, J. van den Hoek, E.J. Dogherty, M. E. Carley, V. Angus, Guideline adaptation and implementation planning: a prospective observational study, Implement. Sci. 8 (2013) 49.

[25] F. Fischer, K. Lange, K. Klose, W. Greiner, A. Kraemer, Barriers and strategies in guideline implementation-a scoping review, Healthcare, Basel, 2016, p. 4.

[26] T.J. Fagan, Letter: Nomogram for Bayes theorem, N. Engl. J. Med. 293 (1975) 257.

[27] S.S. Raab, D.M. Grzybicki, A.R. Hart, S. Kiely, C. Andrew-JaJa, E. Scioscia Jr., Willingness to pay for new Papanicolaou test technologies, Am. J. Clin. Pathol. 117 (2002) 524–533.

[28] A.D. Brown, A.M. Garber, Cost-effectiveness of 3 methods to enhance the sensitivity of papanicolaou testing, JAMA 281 (1999) 347–353.

[29] A.L. Frazier, G.A. Colditz, C.S. Fuchs, K.M. Kuntz, Cost-effectiveness of screening for colorectal cancer in the general population, JAMA 284 (2000) 1954–1961.

[30] A. Gafni, S. Birch, A. Mehrez, Economics, health and health economics: HYES (healthy-years equivalent) versus QALYS (quality-adjusted live-year), J. Health Econ. 12 (1993) 325–339.

[31] D.A. Marshall, B.J. O'Brien, Economic evaluation of diagnostic tests. Vol in: C.P. Price, R.H. Christenson (Eds.), Evidence-Based Laboratory Medicine, AACC, Washington, DC, 2003, pp. 159–186.

[32] H. Sox, The patient-centered outcomes research institute should focus on high-impact problems that can be solved quickly, Health Aff. 31 (2012) 2176–2182.

[33] B. Holmes, G. Scarrow, M. Schellenberg, Translating evidence into practice: the role of health research funders, Implement. Sci. 7 (2012) 39.

[34] M.J. Hallworth, P.L. Epner, C. Ebert, C.R. Fantz, S.A. Faye, T.N. Higgins, et al., Current evidence and future perspectives on the effective practice of patient-centered laboratory medicine, Clin. Chem. 61 (2015) 589–599.

[35] S.E. Kahn, P.M. Jones, A.C. Chin, R.H. Christenson, Defining the path forward: guidance for laboratory medicine guidelines, eJIFCC 26 (2015) 158–167.

[36] D.B. Sacks, M. Arnold, G.L. Bakris, D.E. Bruns, A.R. Horvath, M.S. Kirkman, et al., Executive summary: guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus, Clin. Chem. 57 (2011) 793–798.

[37] M. Zhi, E.L. Ding, J. Theisen-Toupal, J. Whelan, R. Arnaout, The landscape of inappropriate laboratory testing: a 15-year meta-analysis, PLoS One 8 (2013) e78962.

[38] L.H. Hilborne, When less is more for patients in laboratory testing, Am. J. Clin. Pathol. 139 (2013) 271–272.

[39] N.E. Morden, C.H. Colla, T.D. Sequist, M.B. Rosenthal, Choosing wisely--the politics and economics of labeling low-value services, N. Engl. J. Med. 370 (2014) 589–592.

[40] C. van Walraven, C.D. Naylor, Do we know what inappropriate laboratory utilization is? A systematic review of laboratory clinical audits, JAMA 280 (1998) 550–558.

[41] J.H. Nichols, T.S. Kickler, K.L. Dyer, S.K. Humbertson, P.C. Cooper, W.L. Maughan, et al., Clinical outcomes of point-of-care testing in the interventional radiology and invasive cardiology setting, Clin. Chem. 46 (2000) 543–550.

Self-assessment questions

1. The prevalence of pulmonary embolism in the ED population is determined by several studies as being 28%. The positive and negative likelihood ratios of D-dimer are 3.0 and 0.1, respectively. What is the posttest probability that patients with negative D-dimer test result will have a diagnosis of pulmonary embolism?
   a. 4%
   b. 50%
   c. 12%
   d. 7%
2. Which of the following can be used to estimate pre-test/posttest probability?
   a. ROC curve
   b. Fagan nomogram
   c. Forest plots
   d. likelihood ratio plot
3. Which of the following ROC curves indicates an overall excellent diagnostic performance of a test?
   a. A curve that is close to the diagonal
   b. A curve that passes close to the top left-hand corner
   c. A curve that passes close to the bottom right-hand corner
   d. A curve that has a slope of 1
4. Which of the following approaches of health economic evaluation aims to answer the question "given a fixed budget, what is the most efficient way to spend it"?
   a. CBA
   b. CEA
   c. CUA
   d. WTP
5. Which of the following is considered the best evidence for studying a diagnostic test?
   a. a double-blind randomized controlled trial
   b. a SR of observational study
   c. a SR of prospective cohort studies with blind comparison to a gold standard
   d. a prospective observational study
6. Which of the following is consistent with guideline development?
   a. Guidelines should be based on one or more SRs published on the topic.
   b. Because guidelines constantly become outdated, a date must be set to review the guidelines and determine whether an update is needed.
   c. The guideline development group must include members from the clinical and laboratory practice.
   d. All of the above are consistent with guideline development.
7. The diagnostic sensitivity and specificity of a test are 90% and 70%, respectively, at the optimum cutoff. What are the positive and negative likelihood ratios for the test?
   a. positive: 0.22 and negative: 8.0
   b. positive: 8.0 and negative: 0.22
   c. positive: 3.5 and negative: 0.12
   d. positive: 3.0 and negative: 0.14
8. Two tests, A and B, are compared for the screening of left ventricular dysfunction in a primary care clinic. Cutoffs selected show diagnostic sensitivities of 80% for A and 50% for B; diagnostic specificities are 20% for A and 40% for B. The area under the ROC curve is 0.75 for test A and 0.70 for test B. Which test is likely better as a screening test for left ventricular dysfunction?
   a. Test A is likely better.
   b. Test B is likely better.
   c. Test A and B are likely to be equivalent.
   d. There is insufficient information to determine which is likely better.
9. What are the four important factors to consider reducing barriers to clinical practice guidelines?
   a. Awareness, agreement, adoption, adherence
   b. Reduce, reuse, recycle, recover
   c. Ask, acquire, appraise, audit
   d. Accountability, accreditation, assessment, articulation
10. The PICOTS acronym is useful for specifying important aspects of formulating the clinical question and acquiring evidence. PICOTS for:
    a. Paradigm, Index, Comorbidities, Outcome, Timing, Setting
    b. Population, Indicator, Comparator, Outcome, Temperature, Sickness
    c. Paradigm, Indicator, Comparator, Outcome, Timing, Setting
    d. Population, Indicator, Comparator, Outcome, Timing, Setting

Answers

1. a
2. b
3. b
4. b
5. c
6. d
7. d
8. a
9. a
10. d

Chapter 17

Harmonization of results among laboratories

W. Greg Miller
Department of Pathology, Virginia Commonwealth University, Richmond, VA, United States

Learning objectives

After reviewing this chapter, the reader should be able to:

- Explain why harmonized results are important for applying clinical practice guidelines.
- Identify or establish the metrological traceability for calibration of measurement procedures.
- Identify reference materials and reference measurement procedures that are suitable for use to establish or verify metrological traceability.
- Explain why some measurement procedures are not harmonized and advise clinical care providers accordingly.

Why harmonized results are important?

Definition

Harmonization is achieving equivalent results, within clinically meaningful limits, among different measurement procedures for the same laboratory test. Standardization is a closely related term that means equivalent results are achieved by having calibration of measurement procedures traceable to a recognized reference measurement procedure (RMP) or certified reference material (CRM).

Impact of harmonization

Many clinical decisions are based on values from laboratory tests. In many cases, clinical practice guidelines exist that recommend specific medical decisions and actions depending on the value of a laboratory test result. Clinical practice guidelines, whether based on international consensus from outcome data or on recommendations in a textbook, assume the laboratory results are harmonized irrespective of the measurement procedure used or when a measurement is made. Some laboratory tests, such as hemoglobin A1c, cholesterol, glucose, and others, have been successfully harmonized by applying metrological traceability strategies for calibration described in a subsequent section, and are used very effectively in clinical practice guidelines. If laboratory results are not harmonized, then patients may get incorrect treatment based on different results from nonharmonized measurement procedures. Unfortunately, a substantial number of laboratory measurement procedures do not produce harmonized results, and clinical providers have difficulty using such results appropriately.

For example, a recent report showed lack of harmonization for parathyroid hormone (PTH) [1]. Individual patient samples measured by five commercially available PTH measurement procedures had an approximately fourfold difference in results. The impact of these differences was examined using guidelines from the UK Renal Association for treating hypophosphatemia in kidney disease. In this application, the number of patients who were identified as needing treatment varied from one-half of the group to almost all of them based only on which laboratory measurement was used. Clearly, the PTH assays were not adequately harmonized to support this clinical decision. Because of the differences among laboratory test results, the UK Renal Association guidelines recommend making treatment decisions when PTH results are two to nine times the reference interval for an individual measurement procedure [2]. However, the measurement procedures had very similar reference intervals despite differences in the values produced [1,3]. In principle, a reference interval could compensate for lack of harmonization. However, in practice, the difficulty to identify an appropriate reference population and to relate its reference interval to different disease conditions limits the effectiveness of reference interval-based treatment guidelines [3]. Given the limitations for use of PTH results for clinical decisions, a harmonization activity for this analyte

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00017-X

is being pursued by an International Federation for Clinical Chemistry and Laboratory Medicine (IFCC) working group.

Infrastructure to support harmonization

Metrology is the science of measurement. Metrology organizations are typically established by a country or a political unit, such as the European Union. For example, the National Institute for Standards and Technology in the United States and the Joint Research Center in the European Union provide metrology services. Metrology organizations produce CRMs and RMPs. The term metrological traceability refers to a series of calibrations and value transfers from CRMs and/or RMPs to the calibrators used in the clinical laboratory that enables the results from clinical samples to have their calibration traceable to the CRM or RMP that defines the reference system for a particular measurand. The components of a reference system are called "higher order," because they are expected to have lower uncertainty and be an accuracy reference for a clinical laboratory measurement procedure. Prior to 2003, many countries had organizations that approved reference system components as suitable for use in the calibration hierarchies of clinical laboratory measurement procedures.

The European Commission passed a directive that by 2003 all in-vitro diagnostic (IVD) devices sold in the European Union were required to have calibration traceable to a higher order reference material or RMP [4]. The directive was replaced by a regulation in 2017, effective 2022, with essentially the same requirements [5]. In response to the European Commission directive, the International Organization for Standardization (ISO) developed standards for higher order reference system components, and the Joint Committee for Traceability in Laboratory Medicine (JCTLM) was formed to approve reference system components for use by IVD medical device manufacturers. Because most IVD manufacturers sell on a global basis, the ISO standards and JCTLM largely replaced national organizations for the same functions, although some countries still retain calibration standard-setting bodies. In addition, many countries have regulations and a regulatory approval agency such as the Food and Drug Administration in the United States that approve all medical devices including clinical laboratory measurement procedures as safe and effective before they can be sold.

International Organization for Standardization

ISO Technical Committee 212, Clinical Laboratory Testing and In-Vitro Diagnostic Test Systems, has published standards for calibration traceability and the components of a reference system needed for traceability. The applicable standards are:

- 17511:2020, IVD medical devices—requirements for establishing metrological traceability of values assigned to calibrators, trueness control materials, and human samples;
- 21151:2020, IVD medical devices—measurement of quantities in samples of biological origin—requirements for international harmonization protocols establishing metrological traceability of values assigned to calibrators and human samples;
- 15193:2009, IVD medical devices—measurement of quantities in samples of biological origin—requirements for content and presentation of RMPs;
- 15194:2009, IVD medical devices—measurement of quantities in samples of biological origin—requirements for CRMs and the content of supporting documentation; and
- 15195:2018, Laboratory medicine—requirements for the competence of calibration laboratories using RMPs.

Joint Committee for Traceability in Laboratory Medicine

The JCTLM was created in 2002 through a Declaration of Cooperation between the International Committee of Weights and Measures, the IFCC, and the International Laboratory Accreditation Cooperation [6]. The JCTLM reviews and approves reference system components for conformance to the requirements in the ISO standards. Approved reference materials, RMPs, and reference measurement services are listed in the JCTLM database available on their website.

One caution regarding reference materials listed by the JCTLM is that information on commutability (explained in a subsequent section) is not provided. The individual certificate of analysis for a reference material should be reviewed to determine the intended use of a given reference material and if its commutability with clinical samples has been validated. Some pure substance reference materials are intended for use to calibrate higher order RMPs, such as isotope dilution−mass spectrometry, but may not be suitable for use to calibrate clinical laboratory measurement procedures.

How to achieve harmonized results

Two main requirements must be met to achieve harmonized results among a group of clinical laboratory measurement procedures. First, the calibration of all measurement procedures must be traceable to a common reference system. Second, all of the measurement

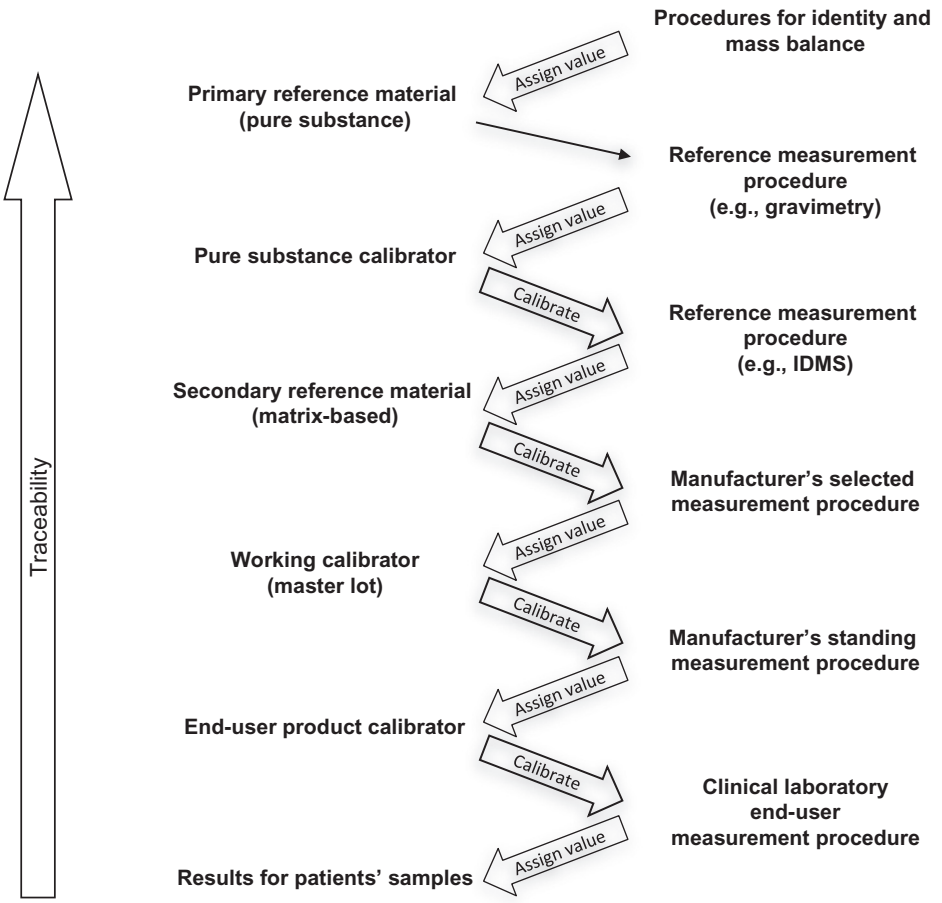


FIGURE 17.1 Calibration hierarchy for metrological traceability of values assigned to patients' samples based on ISO 17511 [8].

procedures must measure the same quantity. For example, if human chorionic gonadotropin (HCG) is to be measured by an immunoassay, then all measurement procedures must respond to the same molecular forms of HCG if harmonized results are to be achieved.

The International Vocabulary of Metrology defines a measurand as the substance intended to be measured that includes the matrix in which the substance is measured and units for reporting the result [7]. Laboratories use the term analyte to refer to the substance being measured. For harmonization of results, the substance actually measured including any molecular forms such as glycosylated, phosphorylated, complexed, or variants, present in disease conditions must be clearly understood. In addition, any other substances in the clinical sample that can influence the measurement must be considered. It is important to clarify a measurement procedure's selectivity for the measurand before attempting to establish its metrological traceability to a reference system. If nonselectivity for the measurand causes substantial differences in results for different patient samples, it may be necessary to improve the selectivity of affected measurement procedures before results can be successfully harmonized.

Metrological traceability of calibration to a reference system

A reference system ideally consists of a pure substance reference material and an RMP calibrated using that reference material. RMPs are generally not suitable for use in clinical laboratories, because they are capable of only making a few very precise and accurate measurements at a time and may require more than 1 day to complete. Measurement procedures used in a clinical laboratory have their calibration traceable to reference systems. Metrological traceability to a reference system can be reproduced over time and location to ensure that calibration of clinical laboratory measurement procedures remains consistent in different laboratories and for tests performed at different times.

Fig. 17.1 shows metrological traceability for the calibration hierarchy of results assigned to patients' samples according to ISO 17511 [8]. At the top of the hierarchy is a pure substance primary CRM that is characterized for identity and purity using appropriate RMPs. A primary reference material has a known mass fraction typically expressed as milligram per gram. For example, a glucose

reference material might be crystalline glucose of 993 mg/g. A CRM is typically provided by a national metrology institute (such as NIST in the United States) or another qualified producer, is well characterized, and is documented with a certificate of analysis. The next component in the hierarchy is a primary calibrator typically prepared by weighing a primary reference material with a calibrated balance (an RMP) to produce a solution of known concentration that is suitable for use as a calibrator for the next RMP in the hierarchy. The International System of Units (Système International, SI) is realized in the pure substance calibrator for example as mol/L. Techniques other than gravimetry are acceptable for preparing a calibrator from a pure substance reference material. Note that a pure substance dissolved in an aqueous solvent may be suitable for calibrating an RMP that is highly selective for the measurand and is not influenced by differences in matrix between the calibrator and other samples such as secondary reference materials or clinical samples that will be measured. However, a pure substance calibrator may not be suitable for use to calibrate some clinical laboratory measurement procedures that are more sensitive to the matrix of the sample being measured.

The next step in the calibration hierarchy is assigning values to a secondary CRM using a properly calibrated RMP. A commonly used measurement technology at this level is isotope dilution−mass spectrometry that is capable of very selectively measuring the measurand with little or no influence from matrix components. An RMP is typically operated with substantial isolation of the measurand from other matrix components, specialized calibration techniques such as calibrators that bracket the concentration in the sample, replication of measurements to reduce uncertainty in the result, and gravimetric preparation of all dilutions in the procedure. Using an isotope dilution−liquid chromatography−mass spectrometry procedure (ID/LC−MS) in a clinical laboratory is not sufficient qualification to be considered an RMP.

For measurands, such as enzymes, the reaction conditions for a measurement procedure determine the catalytic activity and thus define the measurand in SI units as microkatal per liter. In such cases, there is no pure substance reference material, and an RMP is the highest component in the traceability chain. The reaction conditions of the RMP define the SI unit for the measurand and are very completely specified including how to calibrate the equipment and to qualify the reagents used.

For measurands, such as HDL cholesterol or free thyroxine, the RMP defines the measurand, but there are primary reference materials available to calibrate the RMP used to measure the molecule(s) of interest. For example, in the case of HDL cholesterol, the ultracentrifugation conditions determine what fractions of lipoproteins are included in what is called HDL. The measurement of cholesterol is performed by an ID/LC−MS procedure using calibrators prepared from a cholesterol primary reference material. Similarly, for free thyroxine, the equilibrium dialysis conditions determine the amount of thyroxine that is included in the free fraction, but the measurement of thyroxine is by an ID/LC−MS procedure using calibrators prepared from a thyroxine primary reference material. The term international conventional RMP is used for such conditions, because part of the procedure is agreed by international consensus or convention.

A secondary reference material typically has a matrix the same as or similar to that of the clinical samples for which measurements are made. Certified secondary reference materials are produced by a metrology institute or similarly qualified producer and are supplied with certificates of analysis that describe the production of the material, how its measurand value was assigned and a statement regarding the commutability of the material with clinical samples for measurement procedures for which the secondary reference material is intended to be used as a calibrator (see the "Commutability" section). A panel of individual clinical samples may be used as a secondary reference material in the traceability hierarchy. A panel of clinical samples with values assigned by an RMP can be used as calibrators for the next step in the traceability chain. A panel of clinical samples is by definition commutable, because they are the samples intended to be measured and thus can overcome noncommutability limitations of secondary reference materials. However, an individual clinical sample may contain influence quantities (interfering substances) that alter the response of a clinical laboratory measurement procedure to a measurand, and thus make it not suitable for use as a secondary reference material. Experimental designs must take into consideration the possibility of sample-specific influences and include outlier identification as part of the metrological traceability step when individual clinical samples are used as secondary reference materials.

A secondary reference material is used to calibrate a manufacturer's selected measurement procedure. A selected measurement procedure is one that has acceptable performance characteristics for the purpose to assign values to a master lot of calibrator or to an end-user product calibrator intended for use to calibrate a measurement procedure used in a clinical laboratory. An end-user product calibrator is sold along with the instrument and reagents for use in a clinical laboratory for patient sample testing. A selected measurement procedure may be the same as the clinical laboratory measurement procedure but operated more stringently (e.g., with additional replication and/or more stringent calibration) to reduce the uncertainty in values assigned to a master lot of calibrator or to an end-user product calibrator. A master lot of calibrator, called a working calibrator, is

typically used by a manufacturer instead of a more expensive and limited availability secondary CRM to calibrate the selected measurement procedure used in a manufacturing process to value assign multiple lots of the end-user product calibrator over an extended manufacturing time interval. In the case of a laboratory developed test, a selected procedure may be the one that is used for measuring clinical samples. A secondary CRM may be used as a calibrator, or a laboratory may produce and assign values to a master lot of calibrator for daily use to calibrate the laboratory-developed measurement procedure.

The uncertainty of the value assigned to a sample measured at each step in the calibration hierarchy will increase with each measurement in the traceability chain. The uncertainty in the final clinical sample result will be the cumulative propagation of uncertainty at each step. It is acceptable to eliminate one or more steps in the calibration traceability hierarchy when possible, which has the benefit of reducing the uncertainty in the final result for a patient's sample.

Calibration traceability when reference system components are missing

It is desirable to develop pure substance reference materials and RMPs whenever possible, because these components enable a complete metrological traceability chain. Unfortunately, pure substance reference materials and RMPs are only available for a relatively small number of measurands in clinical laboratory testing. The ISO 17511 standard defines several categories of traceability that end at lower positions in the metrological traceability hierarchy.

Traceability to a secondary reference material

For a substantial number of measurands, the traceability hierarchy stops at a secondary reference material. In this situation, the commutability of the secondary reference material with authentic clinical samples is important to ensure the reference material will be suitable for use as a calibrator (see the "Commutability" section).

When there is no RMP, value assignment of the concentration in a secondary reference material can be done in several ways. One approach assigns an arbitrary number of units to the amount of measurand expressed as units per liter. Another approach bases the concentration on the amount of a purified substance added to a base material, for example, a recombinant protein hormone can be added to a base serum pool. The preceding approach allows a nominal concentration to be assigned, but the added substance may differ in reactive characteristics from a native measurand; for example, a recombinant protein will lack glycosylation or other posttranslational modifications and may not react in an immunoassay in the same manner as a native molecule in a clinical sample. An added substance may be of a different molecular form, such as a monomer, than that of the native molecule that may be part of a molecular complex. Another approach to value assignment isolates the measurand from the matrix of the reference material and quantitates it by an alternate analytical technique; for example, a protein can be isolated and measured by amino acid analysis. A common approach is to use a trimmed mean of measurements by a group of clinical laboratory measurement procedures that have been qualified to meet specified performance requirements [9–11].

Whatever approach is used for value assignment, a secondary reference material can be effectively used as a common calibrator even if the actual concentration is not known as long as the assigned value remains stable and the material is commutable with authentic clinical samples for the measurement procedures for which it will be used. In these conditions, harmonization of results can be achieved, and clinical decision values can be established that will be consistent and uniform irrespective of the measurement procedure used.

Traceability to a measurement procedure producer's working calibrator

JCTLM lists primary and/or secondary CRMs and RMPs for approximately 100 measurands. When there is no higher order reference system component available, calibration is traceable to whatever substance a manufacturer or laboratory selects as the best available calibration material. It is a common practice for IVD manufacturers to prepare a master lot of a working calibrator that is used to calibrate the internal (standing) measurement procedure used to assign values to sequential lots of the end-user product calibrator. Although metrological traceability can be described, there is no coordination among different producers of measurement procedures for a given measurand, and, consequently, no uniformity in what substance or matrix is used for calibrators. This category of traceability typically does not produce harmonized results for clinical samples among different measurement procedures.

Traceability to a harmonization protocol

An international conference in 2010 recognized that additional approaches to harmonization were needed to address both the limitation that some secondary reference materials were not commutable with clinical samples and the reality that RMPs and CRMs could not be developed for all of the analytes measured in clinical laboratories [12]. The IFCC Committee for Standardization of

Thyroid Function Tests has pursued an approach to harmonization of thyroid stimulating hormone based on a series of comparisons of results for panels of individual donor sera among clinical laboratory measurement procedures [13]. The manufacturers of the measurement procedures also included their internal master lots of working calibrators in the measurements. The results of each round of measurements were used to assess the status of harmonization, to identify method improvements that were needed, and to develop corrections to each manufacturer's calibration hierarchy to improve harmonization for clinical samples among the different measurement procedures [14,15].

The ISO Technical Committee 212 has developed a standard, ISO 21151:2020, that specifies requirements for calibration traceability to a harmonization protocol. Fig. 17.2 shows conceptually how a harmonization protocol could be designed. Suitable harmonization reference materials are value assigned for the concentration of a measurand. Potentially suitable materials include a panel of individual clinical samples, pools of clinical samples, supplemented clinical samples, or other materials that are commutable with clinical samples (see the next section). The value assignment process can be one of the approaches described earlier for a secondary reference material when no RMP is available or any scientifically acceptable approach. The value-assigned harmonization materials effectively become a type of reference materials that are measured along with working calibrators and end-user product calibrators by all measurement procedures for a given measurand. Based on the results, values are assigned by each manufacturer to their working or end-user product calibrators to achieve harmonization of reported results for clinical samples among all of the different end-user measurement procedures. Such an approach needs to consider how to sustain the harmonization process over time, as new lots of reagents, working calibrators (master lots), and end-user product calibrators are introduced, or new measurement procedures are developed. One approach to sustainability could be to prepare an additional set (or sets) of the harmonization reference materials for use at a future time. If the harmonization reference materials are not stable for storage, replacement panels will need to be prepared. In either case, the harmonization protocol needs to specify how to produce and qualify subsequent preparations of harmonization reference materials to ensure consistent metrological traceability to the protocol. Proficiency testing (PT) or external quality assessment (EQA) using commutable samples provides surveillance of the harmonization status and can identify when adjustments are needed. The harmonization protocol concept is sufficiently general that it can be applied to a range of technical approaches, such as chemical assays, immunoassays, nucleic acid-based testing, etc. The ISO 21151:2020 standard enables a harmonization protocol that conforms to the ISO requirements to be listed by the JCTLM as an acceptable approach for metrological traceability for clinical laboratory measurement procedures.

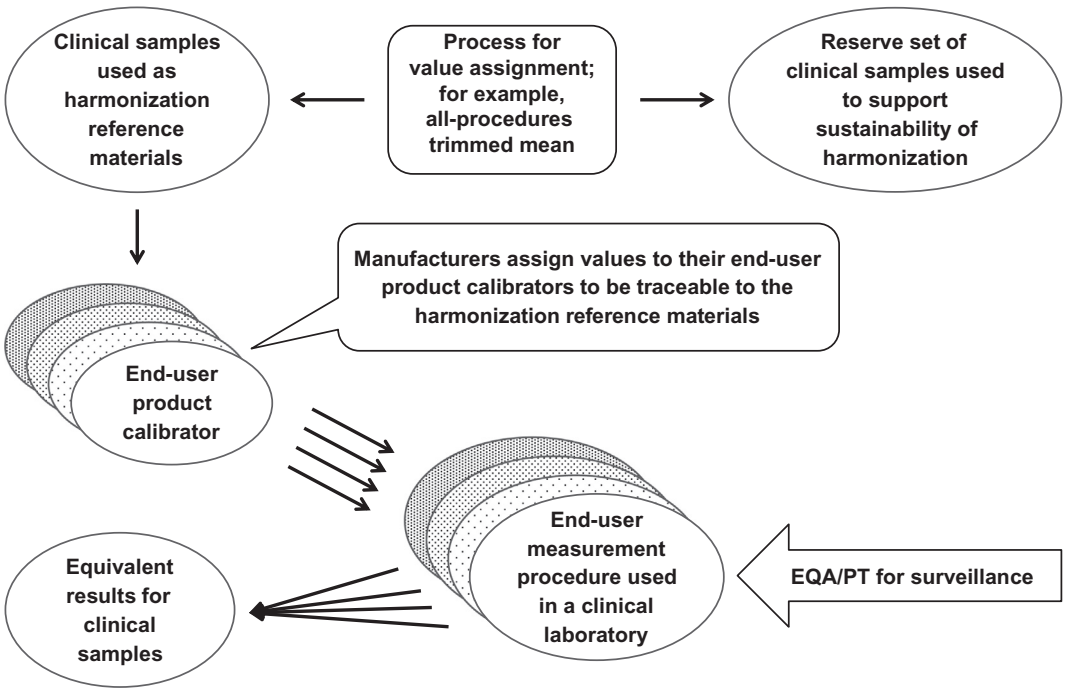


FIGURE 17.2 Example of a harmonization protocol for metrological traceability of calibration to a set of clinical samples used as harmonization reference materials based on ISO 21151.

Commutability

Commutability is defined by the International Vocabulary of Metrology as a property of a reference material, demonstrated by the closeness of agreement between the relation among the measurement results for a stated quantity in this material, obtained according to two given measurement procedures, and the relation obtained among the measurement results for other specified materials [7]. Other specified materials are clinical samples in the case of clinical laboratory testing. Stated differently, commutability is a property of a reference material in which values measured for the reference material and for representative clinical samples have the same relationship between two, or more, measurement procedures for the same measurand [16]. Fig. 17.3A shows the concept of commutability as the same relationship between two measurement procedures observed for reference materials and for clinical samples. Fig. 17.3B shows that noncommutable reference materials have a different relationship between two measurement procedures than observed for clinical samples.

The acceptability of a calibration traceability hierarchy requires that the materials used in each step are commutable with clinical samples among the involved measurement procedures. Because RMPs are designed to be highly selective for the measurand and not influenced by the matrix in which the measurand is found, commutability is generally not an issue for calibrators prepared from primary reference materials, or when measuring manufacturer's working calibrators or clinical samples. Commutability becomes a critical consideration when a secondary matrix-based reference material is used as a calibrator for a manufacturer's internal selected measurement procedure or for a clinical laboratory measurement procedure. Fig. 17.3C shows that using noncommutable reference materials to calibrate each of two measurement procedures causes a difference in agreement for results from clinical samples.

Commutability is a property of a reference material, and correct terminology is to state that a reference material is commutable with clinical samples among a group of measurement procedures. It is incorrect terminology to state that clinical sample results are commutable among a group of measurement procedures. The correct way to state the preceding is that clinical sample results are harmonized, or equivalent, among a group of measurement procedures. Using commutable secondary matrix-based reference materials in a calibration traceability hierarchy is a requirement to achieve harmonized results for clinical samples.

A limitation in laboratory medicine is that a number of secondary matrix-based reference materials intended for use as common calibrators for clinical laboratory

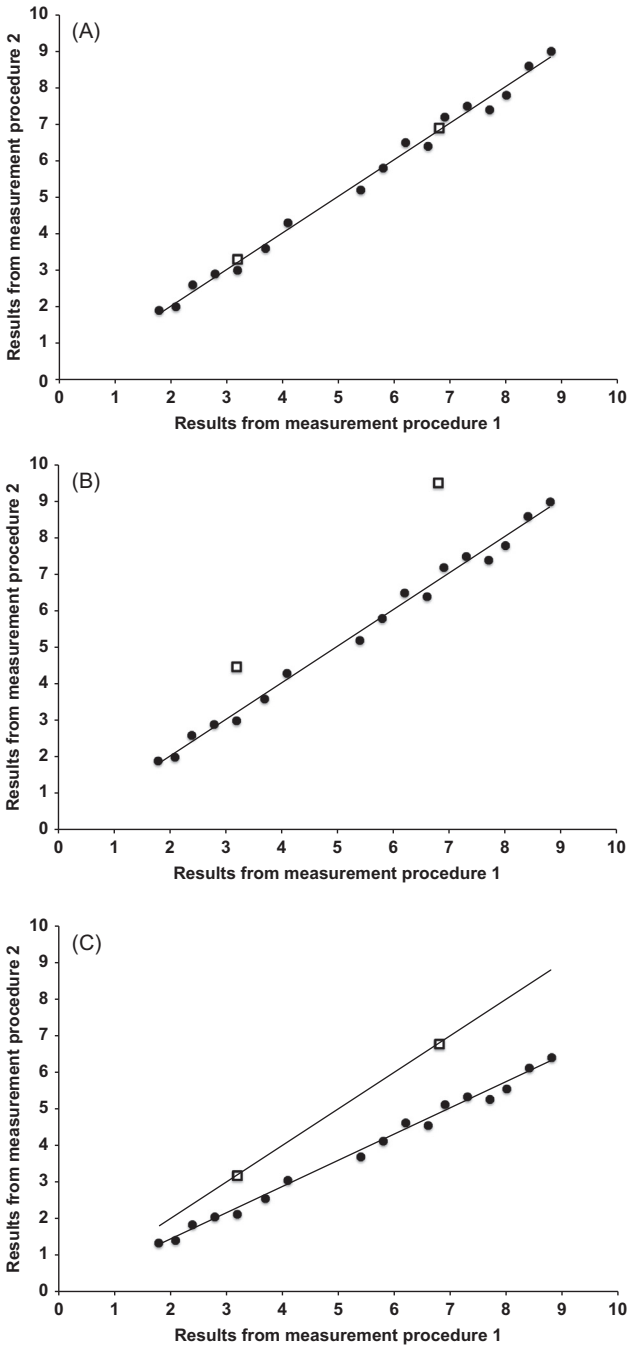


FIGURE 17.3 Plots showing the relationship of results for authentic clinical samples (closed circles) and for reference materials (open squares) measured by each of two measurement procedures. Panel (A) shows that commutable reference materials have a relationship between two measurement procedures that is the same as that observed for clinical samples. Panel (B) shows that noncommutable reference materials have a different relationship than observed for clinical samples. Panel (C) shows that, if the noncommutable reference materials were used as calibrators for each of the measurement procedures, the results for the reference materials appear to be equivalent, but the results for clinical samples are different between the measurement procedures.

measurement procedures have not been evaluated for commutability with clinical samples. A number of investigations have reported that differences among results for clinical samples can be attributed to calibration traceability to noncommutable secondary reference materials [16–18]. Even though the producer of a measurement procedure shows traceability, use of a noncommutable reference material breaks the calibration traceability chain, introduces noncommutability bias, and causes results for clinical samples to be different among different measurement procedures.

Note that most clinical laboratory measurement procedures are not designed to use calibrators prepared from pure substance reference materials in a nonphysiologic matrix, because the difference in matrix composition versus the usual clinical samples makes such calibrators noncommutable. A laboratory that intends to use a pure substance material to prepare a calibrator must validate that calibrator's commutability with clinical samples to ensure the calibrator is suitable for use with the measurement procedure.

Validating commutability of a reference material

Validating commutability of a reference material requires the following steps:

- Obtain representative clinical samples.
- Obtain candidate reference material(s).
- Measure clinical samples and reference materials with all measurement procedures for a measurand.
- Determine the relationships between the measurement procedures for the clinical samples.
- Determine if the relationships for a reference material are close enough to those for the clinical samples for the intended use of the reference material.

If an RMP is available, results from each clinical laboratory measurement procedure are compared with those from the RMP; otherwise, the results for all combinations of two clinical laboratory measurement procedures must be examined.

CLSI has two guidelines for assessing commutability that address many aspects of the experimental design [19,20]. Both guidelines include the common approach for determining the relationship between two measurement procedures by Deming linear regression and calculating the 95% prediction interval statistic. A prediction interval means if another measurement were made, its results would fall within the prediction interval with 95% confidence. A reference material is considered commutable if its result falls within the prediction interval derived from the clinical samples' results. A limitation of the prediction interval and similar approaches is that the prediction interval limits depend on the scatter in results between two measurement procedures. Consequently, each pair of measurement procedures has different acceptance criteria, and those criteria are not related to the intended medical use of results.

An IFCC working group on commutability published three papers in 2018 that provided updated approaches for commutability assessment. The first paper provided general guidance on experimental design [21]. Key design elements include that measurement procedures must meet performance requirements for selectivity for the measurand to be included in a commutability assessment; clinical samples must represent typical pathophysiology but not include unusual specimen specific influences; and the criterion for commutability is fixed for all combinations of measurement procedures and is based on medical requirements for using a test result. The second paper provides an approach for commutability assessment based on the closeness of agreement for the difference in bias between two measurement procedures observed for a reference material compared with the bias observed for clinical samples [22]. A reference material is considered commutable if the difference in bias plus its uncertainty are within a fixed criterion based on the medical requirements for using a laboratory result. The third paper provides an approach for commutability assessment based on substituting a candidate reference material for the current calibrators used in participating measurement procedures' calibration hierarchies, and determining if the improvement in intermeasurement procedure bias meets a fixed criterion based on the medical requirements for using a laboratory result [23]. This latter approach is only applicable to reference materials intended for use as matrix-based secondary calibrators.

How to verify traceability of a clinical laboratory measurement procedure

The manufacturer of a commercially available measurement procedure is responsible for establishing metrological traceability to the highest available level in the calibration hierarchy. The manufacturer is expected to inform the end user of the calibration traceability in the instructions for use provided with the measurement procedure. A clinical laboratory can accept the manufacturer's documentation regarding calibration traceability, and, in most cases, independent verification is not necessary.

When a laboratory wishes to verify calibration traceability, the following approaches can be considered:

- Splitting clinical samples with a laboratory that performs a JCTLM-listed RMP. Unfortunately, there are relatively few such laboratories, and the measurements are expensive. An alternative is to split clinical samples with another laboratory that performs a

measurement procedure similar to that listed by JCTLM with a robust calibration process. Another, less robust, approach is to split clinical samples with another laboratory that performs a measurement procedure that is known and verified to have its calibration traceable to a JCTLM-listed RMP. Results should be interpreted appropriately when using the latter approaches, because the measurement procedures may be less selective for the measurand and have the possibility for a bias in calibration. CLSI document EP09-A3 provides guidance on performing a comparison between measurement procedures using clinical samples [24].

- Measuring a CRM that is documented to be commutable with clinical samples from a metrology institute or other qualified provider. Caution is needed when using some reference materials, because many have not been validated for commutability, in which case they must be treated as noncommutable and so are not suitable for use to determine calibration traceability or bias for a clinical laboratory measurement procedure.
- Participating in a PT/EQA program that uses commutable samples. Some PT/EQA providers offer surveys that use commutable samples, and some also have RMP-assigned values. Results from such surveys can be used to verify the calibration status and bias of a clinical laboratory measurement procedure. Because PT/EQA programs, with some exceptions, typically make a single measurement, the statistical confidence in the results is limited and should not be the sole basis for an estimate of calibration bias. However, the PT/EQA results are a good indicator that calibration traceability is appropriate or may need correction. When commutable samples are used and an RMP value is not available, the results are useful to give a good indication of the harmonization among different measurement procedures. Mean or median values for participants using the same measurement procedures for commutable samples are useful to determine if different measurement procedures are calibrated to give equivalent results and to determine if a measurement procedure is calibrated to be traceable to a reference system when RMP values are assigned.

An important limitation with many PT/EQA programs is that noncommutable samples are used, in which case the results likely contain matrix-related bias and thus are not suitable to estimate calibration bias versus a traceability hierarchy [25]. The same limitation exists for quality control programs that include a comparison of results among different laboratories, because quality control samples are typically noncommutable with authentic clinical samples. These programs are useful to document that a measurement procedure produces results that agree with other laboratories using the same measurement procedure but cannot be used to compare the calibration status or agreement in results among different measurement procedures.

Calibration of laboratory developed tests

A clinical laboratory that develops a measurement procedure is responsible for establishing its calibration traceability. The principles described for IVD manufactures are also applicable to laboratories. Of particular importance is the choice of calibrator. Every effort should be made to use CRMs or international conventional reference materials when available. Attention must be paid to commutability with authentic clinical samples when using matrix-based secondary reference materials as calibrators or for traceability of calibrators prepared by the laboratory.

References

[1] A. Almond, A.R. Ellis, S.W. Walker, Current parathyroid hormone immunoassays do not adequately meet the needs of patients with chronic kidney disease, Ann. Clin. Biochem. 49 (2012) 63–67.

[2] S. Steddon, E. Sharples, Clinical Practice Guideline CKD-Mineral and Bone Disorders (CKD-MBD) Final Version (01/03/2015). Renal Association UK. <https://renal.org/wp-content/uploads/2017/06/ckd-mineral-and-bone-disorders-ckd-mbd204ca231181561659443ff000014d4d8.pdf> (accessed 09.03.20).

[3] C.M. Sturgeon, S.M. Sprague, W. Metcalfe, Variation in parathyroid hormone immunoassay results—a critical governance issue in the management of chronic kidney disease, Nephrol. Dial. Transpl. 26 (2011) 3440–3445.

[4] Directive 98/79/EC of the European Parliament and of the Council of 27 October 1998 on In Vitro Diagnostic Medical Devices, Official Journal of the European Communities L331 (December 7, 1998) 1–37.

[5] Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on In Vitro Diagnostic Medical Devices and Repealing Directive 98/79/EC and Commission Decision 2010/227/EU, Official Journal of the European Union L117 (May 5, 2017) 176–332.

[6] Joint Committee for Traceability in Laboratory Medicine. <https://www.bipm.org/en/worldwide-metrology/jctlm-cooperation/> (accessed 09.03.2020).

[7] The Joint Committee for Guides in Metrology, International Vocabulary of Metrology-Basic and General Concepts and Associated Terms. VIM, third ed., JCGM 200, Sèvres, France, 2008. See <https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf> (accessed 09.03.20).

[8] ISO 17511:2020, In Vitro Diagnostic Medical Devices—Requirements for Establishing Metrological Traceability of Values Assigned to Calibrators, Trueness Control Materials and Human Samples, Geneva, Switzerland, 2020.

[9] I. Zegers, T. Keller, W. Schreiber, J. Sheldon, R. Albertini, S. Blirup-Jensen, et al., Characterization of the new serum protein reference material ERM-DA470k/IFCC: value assignment by immunoassay, Clin. Chem. 56 (2010) 1880–1888.

[10] A. Grubb, S. Blirup-Jensen, V. Lindström, C. Schmidt, H. Althaus, I. Zegers, IFCC Working Group on Standardisation of Cystatin C (WG-SCC). First certified reference material for cystatin C in human serum ERM-DA471/IFCC, Clin. Chem. Lab. Med. 48 (2010) 1619–1621.

[11] S.K. Van Houcke, S. Van Aelst, K. Van Uytfanghe, L.M. Thienpont, Harmonization of immunoassays to the all-procedure trimmed mean - proof of concept by use of data from the insulin standardization project, Clin. Chem. Lab. Med. 51 (2013) e103–e105.

[12] W.G. Miller, G.L. Myers, M.L. Gantzer, S.E. Kahn, E.R. Schönbrunner, L.M. Thienpont, et al., Roadmap for harmonization of clinical laboratory measurement procedures, Clin. Chem. 57 (2011) 1108–1117.

[13] K. Van Uytfanghe, L.A. De Grande, L.M.A. Thienpont, "Step-up" approach for harmonization, Clin. Chim. Acta 432 (2014) 62–67.

[14] L.M. Thienpont, K. Van Uytfanghe, S. Van Houcke, B. Das, J.D. Faix, F. MacKenzie, , et al.IFCC Committee for Standardization of Thyroid Function Tests (C-STFT), A progress report of the IFCC Committee for Standardization of Thyroid Function Tests, Eur. Thyroid. J. 3 (2) (2014) 109–116.

[15] L.M. Thienpont, K. Van Uytfanghe, L.A.C. De Grande, D. Reynders, B. Das, J.D. Faix, et al., Harmonization of serum thyroid-stimulating hormone measurements paves the way for the adoption of a more uniform reference interval, Clin. Chem. 63 (2017) 1642–1652.

[16] W.G. Miller, G.L. Myers, R. Rej, Why commutability matters, Clin. Chem. 52 (2006) 553–554.

[17] W.G. Miller, G.L. Myers, Commutability still matters, Clin. Chem. 59 (2013) 1291–1293.

[18] W.G. Miller, J.R. Tate, J.H. Barth, G.R. Jones, Harmonization: the sample, the measurement and the report, Ann. Lab. Med. 34 (2014) 187–197.

[19] Characterization and Qualification of Commutable Reference Materials for Laboratory Medicine; Approved Guideline, CLSI document EP30-A, Clinical and Laboratory Standards Institute, Wayne, PA, 2010.

[20] Evaluation of Commutability of Processed Samples; Approved Guideline, third ed., CLSI Document EP14-A3, Clinical and Laboratory Standards Institute, Wayne, PA, 2014.

[21] W.G. Miller, H. Schimmel, R. Rej, N. Greenberg, F. Ceriotti, C. Burns, et al. IFCC Working Group on Commutability, IFCC Working Group recommendations for assessing commutability part 1: general experimental design, Clin. Chem. 64 (2018) 447–454.

[22] G. Nilsson, J.R. Budd, N. Greenberg, V. Delatour, R. Rej, M. Panteghini, et al. IFCC Working Group on Commutability, IFCC Working Group recommendations for assessing commutability part 2: using the difference in bias between a reference material and clinical samples, Clin. Chem. 64 (2018) 455–464.

[23] J.R. Budd, C. Weykamp, R. Rej, F. MacKenzie, F. Ceriotti, N. Greenberg, et al. IFCC Working Group on Commutability, IFCC Working Group recommendations for assessing commutability part 3: using the calibration effectiveness of a reference material, Clin. Chem. 64 (2018) 465–474.

[24] Measurement Procedure Comparison and Bias Estimation Using Patient Samples; Approved Guideline, third ed., CLSI Document EP09c, Clinical and Laboratory Standards Institute, Wayne, PA, 2018.

[25] W.G. Miller, G.R.D. Jones, G.L. Horowitz, C. Weykamp, Proficiency testing/external quality assessment: current challenges and future directions, Clin. Chem. 57 (2011) 1670–1680.

Self-assessment questions

1. What is harmonization?
   - **a.** having equivalent results among different measurement procedures
   - **b.** having a suitable reference interval for all measurement procedures
   - **c.** having calibration traceable to a secondary reference material
   - **d.** having calibration traceable to a reference measurement procedure
2. Metrological traceability means _____.
   - **a.** the calibrator value is assigned by a reference measurement procedure.
   - **b.** the accuracy of patient sample results is linked to a reference measurement procedure or a reference material.
   - **c.** the instructions for use describe how a measurement procedure is calibrated.
   - **d.** the lot-to-lot differences in calibrators are less than a specification.
3. What is JCTLM?
   - **a.** an international laboratory accreditation organization
   - **b.** an international group that publishes standards for calibration traceability
   - **c.** an international group that produces reference materials
   - **d.** an international group that approves reference materials that meet requirements specified by the ISO
4. A metrological traceability calibration hierarchy describes _____.
   - **a.** a sequence of steps that transfer values from certified reference materials to the calibrators used in clinical laboratory measurement procedures
   - **b.** a sequence of steps that should be followed when calibrating a clinical laboratory measurement procedure
   - **c.** calibration and quality control procedures for a clinical laboratory measurement procedure
   - **d.** how a manufacturer assigns values to each lot of calibrator for a clinical laboratory measurement procedure
5. What is an essential property for a matrix-based secondary reference material?
   - **a.** commutability with clinical samples
   - **b.** adequate volume to be used for several calibrations
   - **c.** sufficient quantity to be available for at least 3 years
   - **d.** commutability for at least one-half of the commonly used measurement procedures
6. What is a primary reference material?
   - **a.** the best available material to use for calibration
   - **b.** a pure chemical that meets American Chemical Society specifications
   - **c.** a pure substance that is characterized for identity and mass balance
   - **d.** a pure substance that can be added to a pool of clinical samples to prepare a calibrator
7. When there is no primary reference material or reference measurement procedure for a measurand, _____.
   - **a.** a laboratory should use a PT/EQA material for calibration traceability.
   - **b.** a laboratory should use any calibrator as long as a proper reference interval is established.
   - **c.** the calibrator from the manufacturer of a measurement procedure defines the calibration.
   - **d.** a secondary reference material should be the basis for calibration traceability when available.
8. Commutability means _____.
   - **a.** results for patient samples agree among different measurement procedures.
   - **b.** results for PT/EQA samples agree among different measurement procedures.
   - **c.** results for a reference material are the same as the results for clinical samples when measured by different measurement procedures.
   - **d.** the same reference interval can be used for all measurement procedures.
9. A clinical laboratory can verify the calibration traceability of its measurement procedures by _____.
   - **a.** splitting clinical samples with another laboratory performing the same method
   - **b.** measuring a commutable secondary reference material
   - **c.** participating in a PT/EQA survey
   - **d.** comparing its QC results to the interlaboratory mean values among all uses of the same QC material

Answers

1. a
2. b
3. d
4. a
5. a
6. c
7. d
8. c
9. b

Chapter 18

Laboratory information management

Christopher R. McCudden[1], Matthew P.A. Henderson[2] and Brian R. Jackson[3,4]

[1]Eastern Ontario Regional Laboratory Association, The Ottawa Hospital Department of Pathology and Laboratory Medicine, University of Ottawa, Ottawa, ON, Canada, [2]Newborn Screening Ontario, The Children's Hospital of Eastern Ontario Department of Pediatrics, University of Ottawa, Ottawa, ON, Canada, [3]Department of Pathology, University of Utah, Salt Lake City, UT, United States, [4]ARUP Laboratories, Salt Lake City, UT, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Describe the types of information commonly encountered in laboratory medicine.
- Discuss the benefits and limitations of different systems used for information management.
- Describe the basic architecture of LIS.
- Detail the flow of a test result from physician order to reporting.
- Define the nature and use of different messaging standards used in LIS.
- Explain the benefits and challenges of secondary data use.

Introduction

The primary role of a clinical laboratory is to provide information to healthcare providers on the biochemical status of patients; as a result, information management is a central focus of clinical laboratories. A laboratory must manage numerous and diverse forms of data to operate efficiently (Fig. 18.1). Data generated by a lab can take many forms including quantitative results from a chemistry analyzer or a DNA sequencer, image files from a microscope, and free text interpretation associated with a lab result. A large amount of data are collected to track the entire testing process from the sample and order receipt to reporting of final results. This information is often referred to as metadata and amounts to all of the information needed to understand the context of a laboratory result. A large part of the information generated by a laboratory is the metadata associated with a test result. These metadata include what procedures were in use at the time of testing, the time and date of sample collection, the analytical method used for sample analysis, quality control (QC) data, environmental data, and patient demographics.

Reporting a single quantitative clinical laboratory result requires integration of multiple streams of data (Fig. 18.2). Laboratory information systems (LIS) have evolved over time to make this process rapid, traceable, and reproducible. This chapter will describe the core functions of a modern LIS as well as provide supplementary information on management systems commonly found in clinical laboratory settings. The informatics needs of a clinical laboratory include sample tracking, relational database, appropriate QC modules, a document control system, an inventory control system, data exchange interfaces, and process automation. An ideal LIS is a modular software system that integrates all the informatics needs of the laboratory.

The LIS is an essential component of a modern clinical laboratory and fulfills three broad functions within the laboratory: it serves as a relational database, it acts as a data exchange interface, and it operates as a laboratory process automation system. These functions will be described in detail throughout this chapter. Briefly, the database function of the LIS links necessary metadata with a specimen to ensure that results are linked to the correct patient with the correct interpretive aids (e.g., comments, reference intervals, and units). However, this externally reported information represents only the tip of the iceberg when it comes to the data associated with a test result. The laboratory collects and records extensive information internally for quality assurance (QA) and accreditation needs. The LIS relational database organizes and stores all this information so that it can be accessed when needed. As the amount of information stored in a modern LIS increases, so does the value of the information; thus reuse of LIS data for clinical decision support, key performance indicators, and clinical research is common. The second broad role of the LIS is to serve as a hub for data exchange within the laboratory. The LIS is typically part of a larger collection of health information systems, and depending on the degree of information

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00018-1

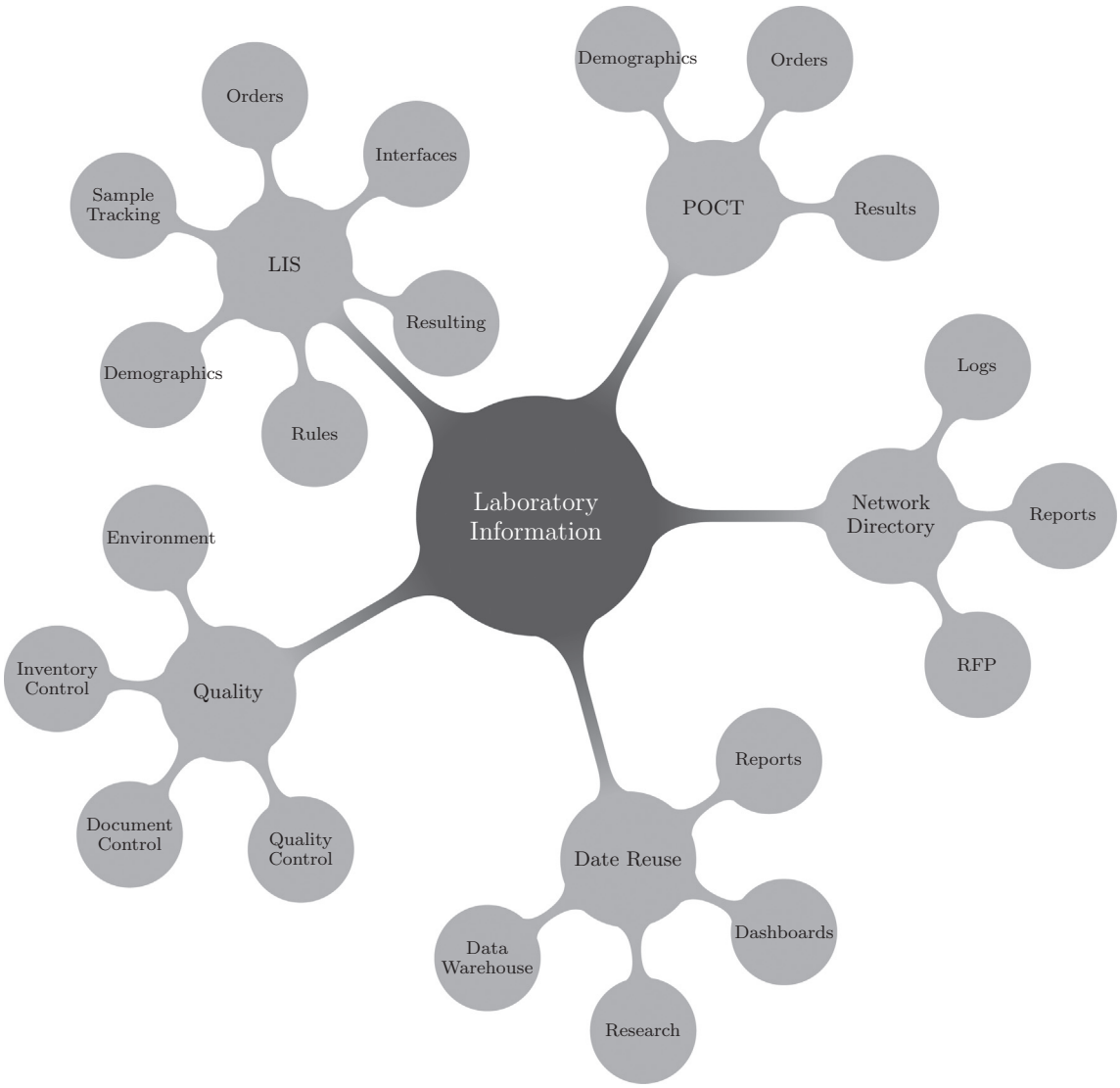


FIGURE 18.1 The informatics requirements of a clinical laboratory. Each secondary node represents a class of functionality required in a clinical laboratory setting. Tertiary nodes show specific examples of functionally.

technology integration that exists within a health system, the LIS can accept demographics and test orders, relay orders to instruments, and accept the verified results. Often, the LIS has interfaces for transmission of data to external systems such as electronic medical records or a data warehouse/repository. Finally, the LIS can automate many aspects within the total testing process; however, the extent of automation is dependent on the functionality of associated hardware (e.g., automation lines). The role of the LIS in process automation can include incorporation of QC rules for autoverification of results, the implementation of laboratory decision support rules such as delta flags, and the appending of prepared comments to results based on established laboratory triggers (or flags).

This chapter will focus primarily on the components of the LIS that are well established in clinical laboratory settings. A fully integrated LIS should allow a laboratorian reliable and timely access to the full breadth of information required to provide high-quality laboratory testing. However, it is important to recognize that laboratory information management is more than the LIS. As additional information relevant to the testing process is commonly not stored in a typical LIS, an overview of information stored outside the LIS will also be included in this chapter.

Management of digital laboratory information

Laboratory information systems

At the heart of clinical laboratory information management is the LIS. The LIS is required at each step of the

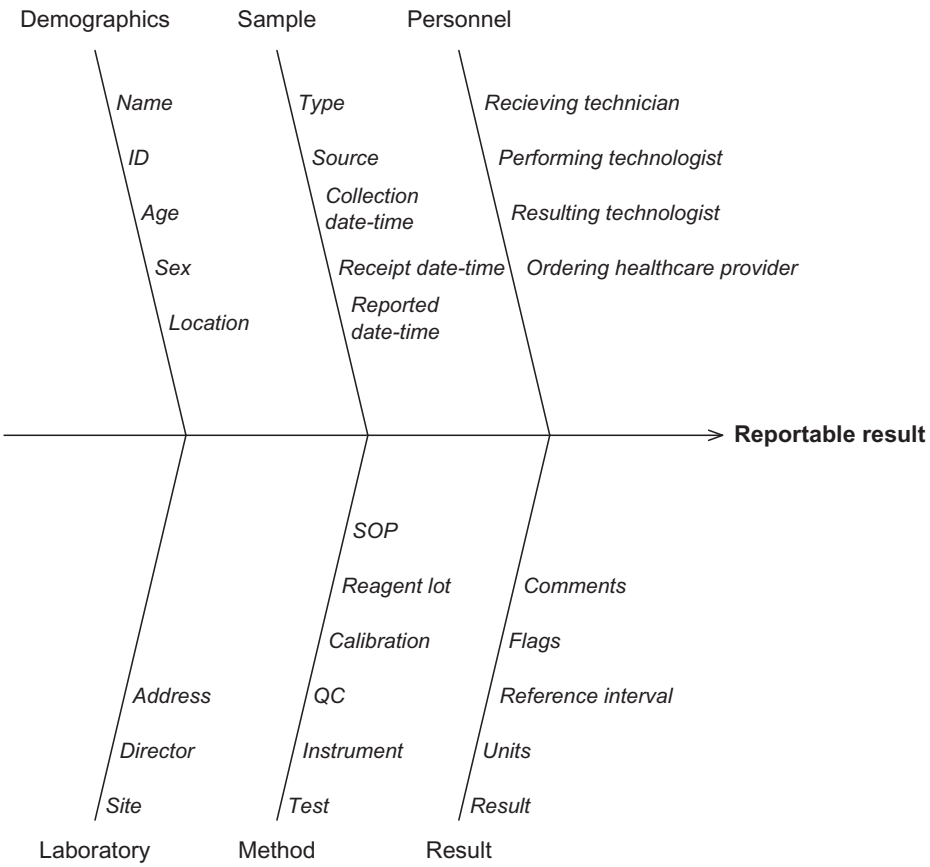


FIGURE 18.2 Streams of data that are linked to a reportable test result.

laboratory testing process, with information flowing to and from the LIS over the course of generating a result (Fig. 18.3). In this section, the testing process is broken down into the familiar preanalytical, analytical, and postanalytical phases for the purposes of illustrating the involvement of the LIS at each step. In the preanalytic phase, the LIS is involved in the processes of test ordering, specimen collection, and specimen processing. During the analytic phase, the LIS is necessary to route the specimen to the appropriate analyzers, to process data received back from the analyzer, and to generate a test result. In the postanalytic phase, the LIS is responsible for result transmission, report generation, and critical value identification and documentation

The LIS is a collection of software, operating systems, and hardware designed to serve the operational processes of the clinical laboratory. The configuration of the LIS is customizable and highly variable, depending on the needs of the individual laboratory. Most available LIS are modular, with separate software options for general laboratory (supporting chemistry, hematology, immunology, and other related laboratory sections), microbiology, and blood banking (also known as transfusion medicine); blood banking, in part due to the dissimilar nature of the information flow and enhanced regulatory requirements, may use a separate blood bank information system, as does Anatomical Pathology. Another key component, which may or may not be a part of a given LIS, is the Admission–Discharge–Transfer (ADT; Table 18.1) system (Fig. 18.3). ADT systems are the backbone of the patient encounter, providing demographic and patient status and location information status used to populate other systems [e.g., LIS, electronic medical record (EMR), and billing]. Underpinning all of these systems are interfaces, dictionaries, and worksheets.

Key concepts

Dictionaries and worksheets

The LIS provides the framework for the laboratory to exert control over the testing process. *The foundation of the LIS is a relational database containing a series of LIS dictionary tables* (Table 18.2). LIS dictionary tables (also known as maintenance tables) define the key elements necessary for the overall operation of the LIS. Examples of items defined in LIS dictionaries include tests, test worksheets, laboratory departments, available specimen and container types, ordering physicians, billing codes, and instruments. In essence, LIS dictionaries contain a dynamic list of all available people, places, and things.

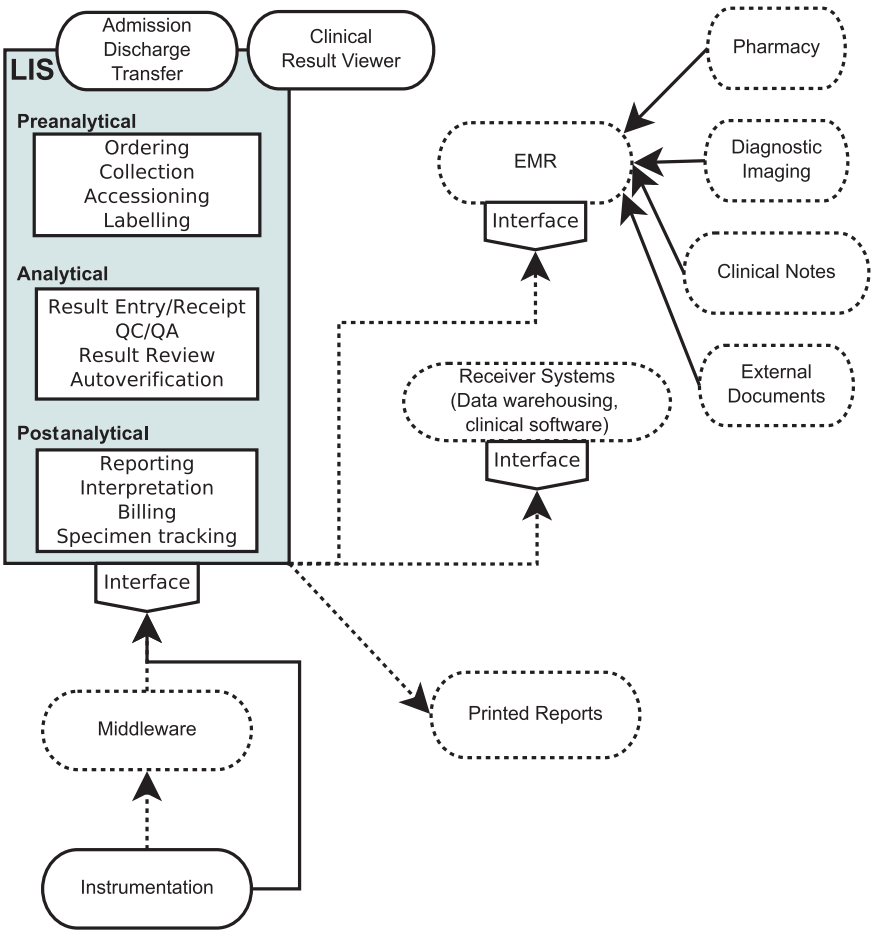


FIGURE 18.3 Overview of laboratory information systems structure and integration. The key components of laboratory information systems are the ability to order tests, collect and label specimens, and report results. Optional external components are surrounded by dotted lines, for example, where results may be transmitted to a data repository. Laboratory information systems have result viewers, though many systems transmit to an electronic medical record. Admission, discharge, and transfer systems also may be external to the laboratory information systems depending on whether the application is a single vendor suite or a composite of several different pieces solutions.

LIS dictionaries populate dropdown lists throughout the LIS application and limit choices in data entry fields to the items in the dictionary list. An example of the utility of this approach can be seen in the manual accessioning process, where the entry of the ordering physician from the ordering physician dictionary list promotes valid data entry; an attempt to order a test for a physician not in the system would be invalid. For each test in the LIS, several interrelated dictionaries define its key elements (Fig. 18.4). Careful attention to the content and maintenance of LIS dictionaries permits the processes of the laboratory to be standardized. LIS dictionaries may be prepopulated by the LIS vendor, but are highly customized to meet the specific needs and workflow of the individual laboratory (this can be both a blessing—in the form of site-specific workflow—and a curse in terms of lack of standardization and interoperability standpoints).

The LIS worksheet (or worklist) is a construct of the LIS that groups tests to be performed on a specific instrument or in a particular laboratory location. The worksheet is an important mechanism for standardizing laboratory workflow, since it links tests, instruments, and QC data. For manual testing, a printable version of the worksheet may be used to organize laboratory workflow and standardize processes. For automated instruments, worksheets are largely "virtual" and specify the elements needed for the correct transfer of data between the LIS and the instruments. Worksheets define the logical flow of a given test and are largely based on the instrument and location. A worksheet for a chemistry instrument, for example, would be linked to the test dictionary, QC and autoverification rules, as well as any associated calculation and available list of order sets.

Laboratory information systems: preanalytic phase

The preanalytical phase (Table 18.3) refers to the parts of the laboratory testing process that occur before a sample is analyzed. The preanalytical phase includes patient identification, sample collection and transportation, test ordering, specimen receipt, accessioning, and container identification (Fig. 18.5). The preanalytical phase accounts for an estimated 40%−70% of laboratory errors. These errors occur largely as a result of manual, highly

TABLE 18.1 Keyterms.

Term	Definition
Accession	Unique identifier for each specimen linked to orders and patient information.
ADT	Admission–discharge–transfer; patient registration system.
API	Application programming interface.
ASTM	American Society for Testing and Materials; defines standards for transmitting information from medical devices.
Big data	Databases that have one or more of the "three Vs": Volume, Variety (particularly of nonnumeric data such as free text), and Velocity (rate of data entering the database).
CID	Container identification; unique number for each tube; may be an extension of accession number.
CPOE	Computerized physician order entry; software allowing physicians to place orders digitally instead of using paper requisitions.
CPT	Current procedural terminology; codes defined by the American Medical Association used for billing; similar to ICD-10, but defines procedures rather than disease.
Discrete data	Quantitative whole number information; for example, counts.
EMPI	Enterprise master patient index; a database that spans a healthcare system, where a single patient identification number can be used across a multisite organization for an individual.
EMR	Electronic medical record; collection of patient health information stored in a digital format. Also called a clinical information system. Electronic health record is sometimes used as a synonym, though this may refer to a collection of EMRs.
Flat file	Plain unformatted text stored in a structure way (e.g., comma separate values).
HIPAA	Health Insurance Portability and Accountability Act; legislation regarding security and privacy of health information.
HIS	Hospital Information System; integrated clinical, financial, and operational information.
HL7	Health Level 7; standard messaging format for transmission of laboratory results.
ICD-10	International Classification of Diseases 10th edition; standard set of terms to catalog medical conditions.
Informatics	Use, storage, and acquisition of digital information to answer questions.
Interface	Computer system (driver) that translates and routes messages from one system to another.
LOINC	Logical observation identifier names and codes; standard naming system for laboratory tests.
Metadata	Information about a file or piece of information, e.g., file size, who created the file, date created.
Relational database	Information formally stored in tables, which use a set of rules and keys to connect to each other.
SNOMED	Systematized nomenclature of medicine; comprehensive set of medical concepts and terms used globally.

TABLE 18.2 Laboratory information systems dictionary fields.

Patient	Physician	Analyte	Sample
ID number	ID number	Name	Location
Date of birth	Name	Result	Encounter
Sex	Specialty	Flags	Reference interval
		Verification date-time	Ordered date-time
		Comments	Received date-time
			Priority

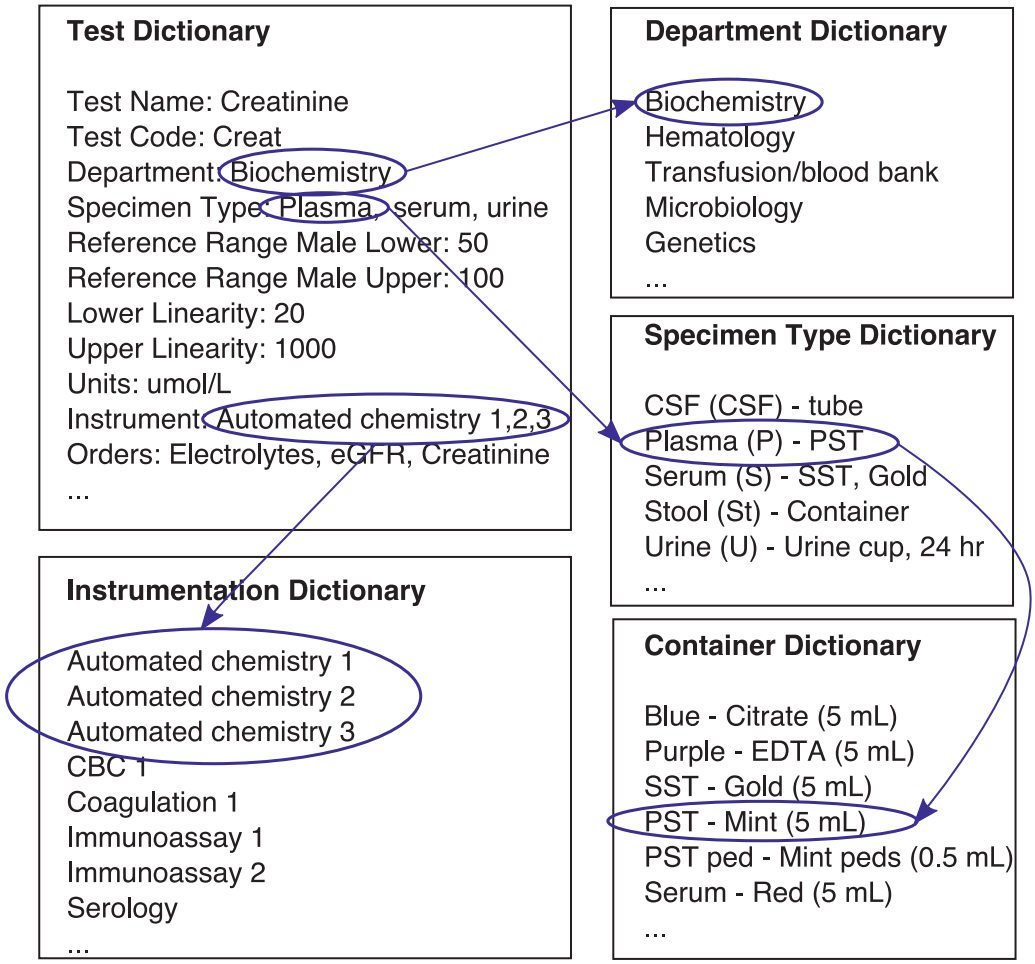


FIGURE 18.4 Laboratory information systems maintenance tables (dictionaries) define the key parameters of each laboratory test.

TABLE 18.3 Laboratory information system function in different testing phases.

Testing phase	LIS involvement in the testing process
Preanalytic	Patient demographics transmitted from registration system to LIS Tests requested by computerized physician order entry or manual test requisitions LIS converts test requests into orders and creates accession number for that order LIS provides collection lists of pending orders to phlebotomy service LIS creates barcoded specimen labels for tube labeling LIS electronically transmits orders to analyzer performing test LIS directs specimen for processing and analysis
Analytic	LIS creates instrument worklists LIS stores quality control data LIS verifies that results fall within reportable range and no flags requiring attention have been generated LIS autoverification finalizes result or a technologist reviews and releases the final result
Postanalytic	LIS sends electronic message with final result to clinical information system If test result is a critical value, the LIS alerts the laboratory staff to contact care unit emergently Clinical staff interpret results and act LIS sends test information to billing system

LIS, Laboratory information system.

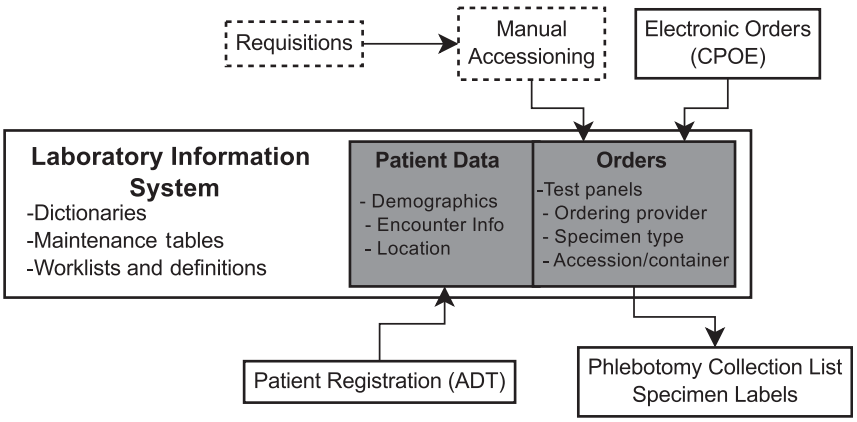


FIGURE 18.5 The role of the laboratory information system in the preanalytic phase of laboratory testing. The dashed boxes indicate error-prone processes.

variable processes, which are outside of the immediate control of the laboratory and include paper requisitions, manual registration, and manual LIS data entry. Preanalytical errors can be prevented with proper deployment of the LIS. The LIS can facilitate quality and safety by providing electronic interfaces for registration and ordering, phlebotomy collection lists, barcoded specimen identification, and process standardization.

Patient identification

Medical records rely on unique and permanent patient identifiers. Patient-specific information (e.g., a combination of name, date of birth, and health ID number) must be entered into the system before tests can be requested. In most health systems, the registration process involves acquiring patient demographic information. Patient demographics include patient name, date of birth, address, phone number, race (and sometimes religion), and sex. An important emerging LIS challenge is how to document nonbinary sex and gender. Briefly, demographic sex is defined as the biological assignment at birth based on external genitalia and chromosomes, whereas gender is a social/cultural construct and personal sense of being. While most LIS are designed to document binary sex only, an appropriate reference interval is required for result interpretation. Currently, there is a paucity of reference interval data on transgender patients, particularly those transitioning and on gender affirming hormone therapies. Consequently, laboratories should consider how to report patient results, either with cautionary notes or references to the albeit limited literature. Discussion with physicians is helpful in instances of hormonal therapy where the ability to provide appropriate reference intervals is limited. The situation is often complicated by the inclusion of biological sex in rules, calculations, and billing systems. Laboratories require procedures and practices to address these issues in the interest of patient care. In support of removing barriers to gender diverse patients, some hospital information systems (e.g., Epic) provide both preferred names and legal names and two- or three-step gender questions to document desired gender identity [i.e., female, male, transgender female (male to female), transgender male (female to male), other, or gender neutral]. Collectively, this is a complex topic with further guidance and software developments needed.

Using the aforementioned demographics, the registration system assigns the patient a unique medical record number (or retrieves it from a database) and electronically transmits patient demographic information to downstream systems, such as pharmacy information systems, clinical information systems, and the LIS. The registration system (ADT) transmits this information via a specially formatted message that is readily interpretable by the downstream systems; ADT messages conform to standards known as Health Level 7 (HL7). Where there are multiple institutions in a system, the entire health system may use an enterprise master patient index (EMPI) number. In such instances, the EMPI is used by all clinical systems at all institutions with a system, even if they also use a separate institutional patient identifier. The EMPI number allows the enterprise to maintain demographics centrally, simplifying the creation of a unified, enterprise-wide electronic patient record. Demographic information can also be manually entered into the LIS for patients that are unknown to the health system, as may occur with specimens sent from laboratory outreach clients. Irrespective of the scale and entry method, the LIS relies on patient identifiers and demographic information to report results.

Order generation

Orders are entered into the LIS either manually through staff who record paper requisitions or electronically in the form of computerized physician order entry (CPOE).

CPOE systems facilitate effective utilization practices by providing clinical order sets and care pathways based on the patient's clinical state. Since the physician is able to select only from available orderable tests, electronic order entry enables the removal of obsolete tests and the introduction of more appropriate options. CPOE applications have the additional potential for clinical decision support, such as test usage and reflex testing algorithms or suggestions for best practices. In CPOE systems, orders are transmitted electronically to the LIS at the time the provider requests the test. The LIS has order entry functionality where test requests, whether arriving via electronic messages or on paper requisitions, are converted into laboratory orders. In the case of paper requisitions, laboratory staff must manually enter the laboratory orders in the LIS upon receipt of the specimen and test requisition by the laboratory. The orders themselves come from physicians or other authorized ordering providers (nurse practitioners, midwives, and pharmacists depending on the state, province, and federal regulations). From workflow and error reduction perspectives, orders should come directly from the provider, whether electronic or paper-based, since misinterpretation and errors may occur when intermediates are involved in the order entry process.

Specimen collection

Specimen collection is driven either by batch collection lists from the LIS (e.g., morning collection rounds) or by *ad hoc* based on individual orders (e.g., outpatient collection or patient bedside). Typically, in a hospital system, collection lists drive the majority of inpatient collections with the goal of providing timely results for morning/evening rounding while taking advantage of the skills of phlebotomists. For collection lists, the LIS groups the orders by the desired collection time. Prior to the desired collection time, the LIS prints specimen labels in batches that contain at least two patient identifiers and specimen type (e.g., tube type) to collect. For *ad hoc* orders, specimen labels are produced on demand at the patient bedside with the use of specialized software and portable devices that permit scanning of a patient's wristband barcode and real-time printing of specimen labels. These are also known as "just-in-time" barcoding, where electronic orders are downloaded to portable devices, which print the appropriate labels and provide direction as to the type and number of containers needed to fulfill the order. Examples of devices used to scan wristbands and generate labels for blood tubes include Rover (Epic) and Bridge (Cerner; some versions). These systems close the loop on patient identification and eliminate many of the issues that arise by printing labels remote from the patient. For outpatient collections, labels and collection details (how many of each type of tube) are printed on demand based on the order. In all instances, samples are labeled with order, test, and demographic information and sent to the laboratory for analysis.

Specimen accessioning

The time at which a patient's orders are placed into the LIS, whether by electronic transmission from an order entry system or direct entry by laboratory staff, is known as the accessioning point. At the time of accessioning, the specimen receives a unique accession number (sometimes called a specimen number) used to track the specimen throughout the testing process. The accession number links the test orders and patient information to the specimen. In order to keep the accession number short, some LIS systems recycle accession numbers. Other LIS systems append the date to the accession number or utilize some other method to keep the accession numbers unique and permanent. In most LIS systems, each part of a given specimen (e.g., three different tubes that are drawn for one patient) is identified by a container ID (CID). The CID is a unique number that distinctly identifies every tube or container in a given specimen. If additional aliquots are obtained from the original tubes, those additional aliquot tubes are typically identified with unique CIDs. The CID is the key identifier that links order and result data in the analyzer and the LIS. The use of unique CIDs is a requirement for the use of robotic automation systems and is essential for specimen retrieval for adding on tests after the original order.

An example of the accessioning process is as follows: a provider fills out a paper requisition, requesting a basic metabolic panel (BMP) and a complete blood count (CBC). Two tubes (one lithium heparin tube for the BMP and one potassium EDTA tube for the CBC) are received by the laboratory along with a paper test requisition. At this point, the laboratory staff accessions the specimen and the LIS assigns an accession number to the entire order. The two individual tubes are assigned unique container identifiers, and the test orders are associated with those CIDs. The LIS generates tube labels that include the CID barcoded on the label.

Other information typically entered into the LIS at the time of test accessioning includes visit information for the patient-provider encounter that resulted in the laboratory testing. This visit information (often a unique "visit ID" or laboratory encounter ID) is stored in the LIS and is associated with the specimen. The visit information is important for patient billing, since the linkage of the laboratory testing to the clinical encounter may be needed to justify the medical necessity of the testing. Other data elements captured during the accessioning process that may be important for billing and reimbursement may include

the patient's diagnosis code [International Classification of Diseases, 10th Revision (ICD-10 code)] for that visit.

Specimen identification

Specimen identification is essential for the accurate reporting of results. Manual transcription, as may occur during manual laboratory accessioning, is likely to be the most important source of laboratory errors, with one in every 300 keystrokes estimated to be in error. Transpositions of digits or misplaced decimal places are especially difficult to detect as errors and can lead to result misreporting. In the preanalytic phase, CPOE and electronic transmission of orders to the LIS are important strategies for reducing manual transcription and its attendant errors. In addition, barcodes are heavily utilized in the laboratory to prevent transcription errors, as well as to identify and track specimens throughout the testing process. In most systems, each tube of a specimen is uniquely identified with a barcode. Barcode specimen labels allow simplified specimen tracking throughout the laboratory with the use of automated barcode scanners or manual scanning by laboratory staff. Most highly automated instruments will have barcode scanners incorporated into the instrument to permit automatic identification and tracking of barcoded specimens.

For tube identification, laboratories typically use linear or 1-dimensional barcodes (a series of black lines and spaces). The linear barcode on the LIS specimen label typically encodes the CID. Formats for the linear barcodes used in laboratories include Code 128 and Code 39. A limitation of linear barcodes is that they can be subject to reading error if the print quality is low or the label is poorly applied. Two-dimensional barcode formats, such as PDF417 and Data Matrix, store data using two dimensions of a surface. These formats enable more data to be represented in the barcode, with some 2-D formats holding hundreds of characters in a single square centimeter. The major disadvantage of the 2-D barcode formats is that they need a more sophisticated barcode reader (scanner) than the simple readers used for linear barcodes.

Radiofrequency identification (RFID) is a newer autoidentification technology that utilizes radio waves to identify objects. RFID systems utilize a small electronic tag that is read by an RFID scanner analogous to the way that a barcode scanner reads a printed barcode. The RFID tag consists of a small integrated circuit attached to an antenna. Data regarding the tagged object is stored in the memory of the integrated circuit. Information is sent wirelessly to or from the RFID tag via radiofrequency signals generated by the RFID scanner. RFID tags can hold large amounts of information, can be very rugged, and can be read through other materials, not requiring a line of sight between the reader and the tag. RFID tags may be well suited for numerous applications in the laboratory testing process including specimen tracking and patient wristband identification. Several automation vendors are currently using RFID technology as the basis of their automation specimen tracking systems. While they are used for carriers on automated lines, RFIDs are considered too expensive (in the current landscape) for widespread implementation and replacement of labels and barcodes.

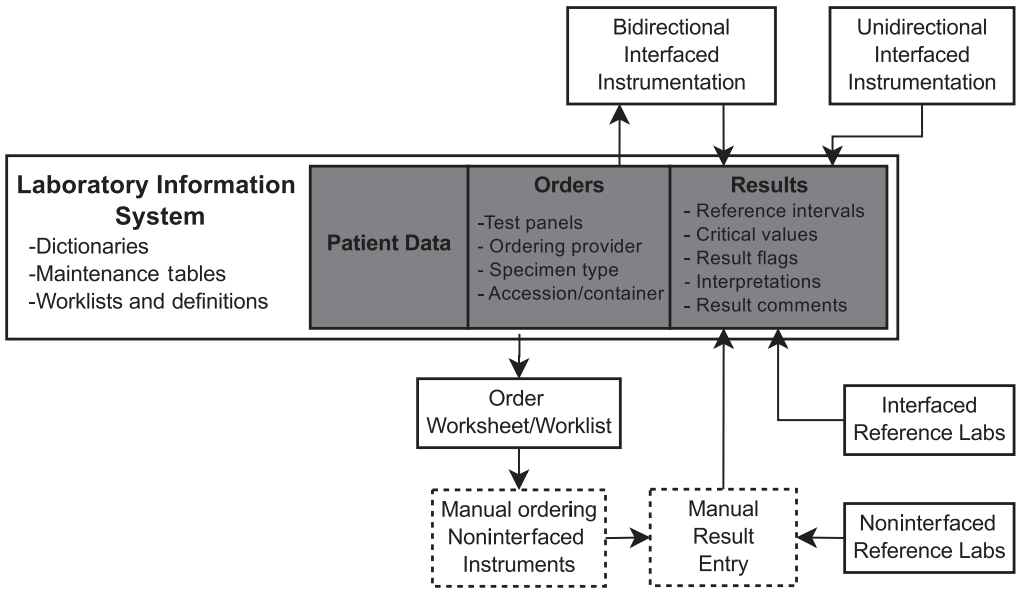


FIGURE 18.6 The role of the laboratory information system in the analytic phase of laboratory testing. The dashed boxes indicate error-prone processes.

Laboratory information systems: analytic phase

Manual results entry

Results are manually entered into the LIS when instruments are not interfaced (Fig. 18.6). These instruments may include osmometers, spectrophotometers, readers (e.g., pregnancy tests), and other technologies that are utilized too infrequently or are performed manually, making interfacing less advantageous from an economic perspective. For low-volume testing platforms, LIS interfaces can be costly in terms of both the capital and labor to install and validate the interface. For testing performed on noninterfaced instruments, the LIS typically provides a printed worksheet of pending orders organized with blank space to permit manual result transcription and to facilitate subsequent manual result entry in the LIS. Given the high rate of transcription errors due to keystroke mistakes, it is essential to have a QA plan in place to oversee manual resulting. One example of such error detection measures is the use of delta checks to identify mislabeling. Another option is for a secondary analyst check the transcription and result entry. In some jurisdictions, secondary checks of manually transcribed results are a regulatory requirement (Institute for Quality Management in Healthcare).

Instrument interfaces

Most laboratory tests are performed on analyzers that exchange data with the LIS through instrument interfaces. Interfacing high-volume automated instruments with the LIS greatly improves productivity and reduces errors. To interface an instrument, the LIS must be configured with software called a device driver or medical device interface (MDI) that allows the LIS to communicate with the instrument. Instrument manufacturers typically provide interface specifications to LIS vendors in advance of instrument introduction. Most instrument−LIS interfaces are based on standards and protocols defined by the American Society for Testing and Materials (ASTM), though newer standards that have been actively developed over the past 5 years are likely to supplant and improve current interoperability (e.g., Laboratory Analytical Workflow standards).

Instrument interfaces can be unidirectional or bidirectional. Unidirectional interfaces solely send results from the analyzer to the LIS. This type of data transfer may also be referred to as a data "upload," the direct transfer of results from the analyzer to the LIS. In a unidirectional interface, there are no data transmitted from the LIS to the instrument, so orders and patient/specimen-identifying information must be entered into the analyzer. The unique CID for each specimen container can serve as the identifying link in these scenarios, since the CID in the LIS is associated with the patient demographics. In addition, since the CID is typically barcoded on the specimen tube, having the analyzer scan the tube barcode can provide a simple way to provide specimen-identifying information to the analyzer. Laboratory blood gas instruments are often configured with unidirectional interfaces, since the full blood gas menu is performed on every sample, and thus the LIS to analyzer transfer of orders is not necessary.

A bidirectional interface permits the LIS and the analyzer to send data back and forth, eliminating the need for the direct entry of orders and patient/specimen identifiers in the analyzer. Bidirectional interfaces are typically required for random access analyzers (such as chemistry systems) where a wide variety of different assays may be performed on a given sample. With a bidirectional interface, the LIS sends patient/specimen identifiers and orders to the analyzer and the analyzer sends the results to the LIS. If the laboratory is already entering orders into the LIS, a bidirectional interface eliminates the need for reentering orders and patient/specimen identifiers in the analyzer. The transfer of patient/specimen identifiers and test orders from the LIS to the instrument is referred to as a "download" (the instrument downloads the information from the LIS). Depending on the sophistication of the interface and the instrument, bidirectional interfaces may send orders/demographics to the instruments in batches (batch download mode) or may dynamically send the information to instruments when requested (host−query interface mode).

Quality control and quality assurance

The LIS is essential to the laboratory for QA. The LIS provides the data elements necessary for the monitoring of assay turnaround time (TAT), since it timestamps the key events for a laboratory test: ordering time, collection time, in-laboratory receipt time, accessioning time (time orders are entered into the LIS), and result time. Specimens that have been received by the LIS but not yet resulted will appear on the technologist's pending test list, compelling them to investigate specimens that have remained on the pending test list for longer than the usual periods of time. In addition to TAT reports, a wide variety of other management reports are available from the LIS, including testing volumes, workload statistics, and billing. Further, the ability to create "*ad hoc*" reports from the LIS is supported in most LIS platforms. A sample *ad hoc* report would be to "select all calcium results from the last 12 hours that were below the reference interval"; further specific examples are shown later in this chapter. The LIS should provide views and reporting tools such that staff can easily create reports to ask important management and clinical questions. As

more data are available, an evolving role for laboratorians is to extract, analyze, and interpret information from the LIS to add value to the testing process.

Results review

The validation of results is an important function of the LIS. The preliminary result of an assay is flagged as "preliminary" or "performed" as it awaits technologist review. The LIS then provides an interface where technologists can review and release (e.g., finalize or verify) results. To ensure that only valid results are released, the LIS may apply rules to facilitate comparison of results against other laboratory metrics, including the reportable range for the analyte, the analyte reference range, or critical value limits. Critical value limits are defined in the LIS test dictionary. Critical values are defined as test values that would place the patient in immediate danger if not promptly treated. An example of a critical value would be a potassium result of 8.0 mEq/L. Such a patient with severe hyperkalemia would be at risk of potentially fatal abnormal heart rhythms. It is imperative that the LIS immediately alert a technologist to a critical value, since the patient's care team needs to be contacted immediately. Many LIS have modules that alert laboratory staff to the presence of a critical value result and permit documentation of the communication to the care team, both of which are necessary for patient care and regulatory requirements. In addition, the current results may be compared to the patient's prior results (a delta check) to alert the technologist to possible erroneous results. Delta checks may alert the technologist to a possible patient identification error or the change in test results may reflect a physiologic event in the patient.

Rules and calculations necessary for result reporting also typically reside in the LIS. Calculation of the international normalized ratio, the anion gap, and the estimated glomerular filtration rate are examples of calculations performed by the LIS. It is essential that the laboratory carefully tests calculations, especially in situations involving unknown patient age or sex, out of range results, and corrected results. Rules in the LIS also provide support for laboratory reflex algorithms. An example of a rules-based reflex algorithm would be to perform testing for free T4 on a specimen where the thyroid-stimulating hormone level was elevated. Reflex algorithms require the LIS to generate automatically orders for the additional testing based on the initial result.

Autoverification

Traditional technologist result verification, as described above, relies on "mental algorithms" that are applied to single results or groups of results. Reliance on individual judgment can lead to variability in process and error. Autoverification (automatic verification or autovalidation) refers to the process where computer-based algorithms perform actions on laboratory results without technologist review. The automated actions may include releasing results to the EMR, sample dilution, cancellation of associated orders, repeat analysis, reflexive testing, or adding comments to specific results. Most commonly, autoverification algorithms are utilized to release automatically results without the need for technologist review. To facilitate autoverification, most LIS have a parameter called the "release range," which can be defined for each test expressly for the purpose of autoverification. A carefully designed autoverification algorithm can lead to improved TATs, increased operational efficiency, and a more consistent process. In particular, it is not possible for humans to review thousands of results over a shift with the attention to detail required to identify potential errors or subtle patterns.

Autoverification rulesets may reside in the LIS or in middleware (Fig. 18.3). In order for autoverification to be implemented safely, the characteristics of the test must be thoroughly validated, and algorithms must be designed into the LIS. A typical autoverification rule would permit test values for a particular analyte to be autoreleased, provided the resulted value falls within a specified autoverification range (e.g., the release range), no QC failures are present, and no instrument flags are present. Delta checks are also commonly incorporated in autoverification algorithms. Autoverification algorithms should be thoroughly tested and periodically reevaluated to ensure that the rules are performing as expected. There are regulatory requirements (e.g., College of American Pathologists) for policies, audit trails, and autoverification retesting at least annually and each time there are system changes. Specific requirements vary with the regulatory body, but may include the need for procedures and testing of QC rules, dilutions, repeat testing, flags, and limits.

Middleware

In addition to test results, decisions during the analytic phase of laboratory testing may depend on clinical data such as the medication list, unreported test parameters, additional analyzer data, QA/QC data, the results of other tests, as well as historical test performance data. Many current LIS platforms do not have access to these data and alone have the capacity to generate an integrated display of such disparate information or to incorporate these data elements into rules and algorithms. To fill this need, laboratory "middleware" [software in the "middle" among the LIS, laboratory instruments, or point-of-care (POC) devices] has been developed to extend the capabilities of the traditional LIS. Middleware may be provided by the

instrument vendor or a third party and provides an environment for sophisticated rule-based processing.

The more advanced rule-based approaches, available through middleware and some advanced LIS vendors, allow rules to be authored that can provide control of retesting, alert when interfering conditions are present, permit reflex testing across multiple specimens and instruments, as well as support automated test add-ons and specimen tracking. In addition to its use for rules-based processing, middleware is often required for the high-resolution specimen tracking necessary for robotic automation lines and for centralized control of instrumentation.

Laboratory information systems: postanalytic phase

Reporting

The postanalytic phase of testing involves result access and interpretation by clinicians (Fig. 18.7). Optimal reporting relies not only on data availability but also on a presentation that makes result interpretation intuitive, efficient, and unambiguous. Further, laboratory reports may serve important medicolegal purposes. Well-designed electronic reporting applications can improve retrieval efficiency and clarity by highlighting recent results, grouping related tests, and enabling clinicians to filter results based on time intervals. Moreover, the availability of customizable, user-friendly charts and graphs allows identification of time-based patterns and intertest relationships. The laboratory does not typically control data displayed within the results viewer such as the hospital's EMR system or personal digital assistant-type results viewers. However, laboratory accreditation requirements require that the laboratory validates the display of laboratory results in downstream result-viewing applications. The laboratory review of result display should pay particular attention to the appearance of result flagging, reference intervals, test and specimen comments, and corrected results. In addition, the laboratory must ensure that the format of results printed out of the EMR and other applications (including the LIS itself) are unambiguous and meet all regulatory requirements for printed laboratory reports.

Result interpretation

With the increasing number of test options and complexity, many clinicians are unable to maintain sufficient knowledge to interpret complex laboratory test results in all clinical circumstances. Rules can be implemented in the LIS to attach coded comments (short interpretive paragraphs) to provide interpretive support for complex test results. For example, hepatitis B test panel results may be accompanied by an automatically generated coded comment that can differ based on the results of the hepatitis panel. In recent years, there has been interest in pathologist-provided interpretive reporting for complex laboratory evaluations such as coagulation and immunology workups. In this approach, a pathologist expert reviews the laboratory testing in the context of the patient's clinical condition and creates a patient-specific interpretive report that is transmitted to the medical record and is available alongside the test results. Interpretive laboratory reports may include a narrative describing the differential diagnosis, pertinent details about the assays that may affect their interpretation, and suggestions for additional testing. In several studies, narrative interpretations created by clinical pathology experts appear to provide positive benefits to the physician, the patient, and the hospital.

Billing

The LIS must report to financial systems an accurate accounting of all tests performed by the laboratory. For each transaction, the LIS typically reports to the billing system the test performed, patient demographics, ordering physician, visit information, and patient's diagnosis. Failure to collect and report this information accurately can lead to nonpayment or regulatory noncompliance. The LIS billing interface must be able to manage tests

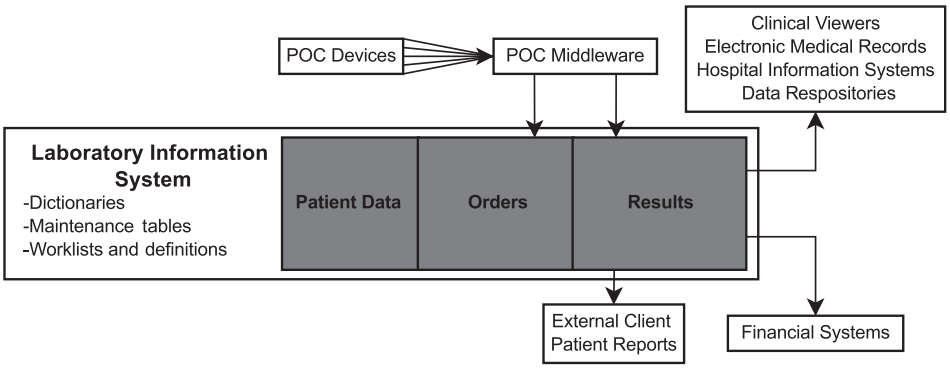


FIGURE 18.7 The role of the laboratory information system in the postanalytic phase of laboratory testing.

added onto the original order and ordered tests that were not performed. The interface with the billing system may dynamically transmit data for tests as they are resulted or the interface may operate in batch mode, sending a daily batch of qualifying tests.

Point-of-care results

With the rapid growth of POC testing, there is increasing interest in having POC results available in the EMR. In most settings, transmission of POC results to the LIS is a requisite step for POC result availability in the EMR. Transmission of results from the POC devices to the LIS occurs via an intermediate data management system (a "data manager" or middleware system) that is interfaced with the individual POC devices. The data manager then transmits data to the LIS. Some data management systems also provide functionality that allows users to result non-interfaced, manual testing (e.g., rapid flu antigen, stool occult blood) into the data manager, facilitating transmission of these results into the LIS and EMR.

POC interfaces differ from other result reporting interfaces, since for POC tests, the interface must simultaneously order and result in testing. Recent standardization of the interfaces among the POC devices, data managers, and LIS has improved the transmission of POC results to the LIS. In addition, this standardization has enabled the development of "vendor neutral" data managers capable of interacting with a wide array of POC devices from different manufacturers. This permits an organization to have a single interface between the data manager and the LIS transfer a wide array of POC test results.

Data exchange and standards

LIS were once largely standalone systems where laboratories managed data generated within the laboratory. This is no longer the case with the LIS being highly connected and involved in all phases of the diagnostic testing process, from test ordering to interpretation. The current generations of LIS are an integral part of the healthcare system and must interact with numerous other systems to function effectively (Fig. 18.3). The LIS must exchange data with practice management systems, hospital information systems, electronic medical record systems, and billing systems.

The connection of the laboratory to the rest of the hospital relies on standardized message formats and communications. Data exchange between these various information systems occurs via application interfaces. These interfaces include the software, protocols, and connections that enable electronic data exchange from one system to another. Common messaging formats and communication protocols serve to standardize data exchange between applications. The recent development of interface engines, which serve as "traffic cops" to route data between applications, has also improved communication in healthcare organizations.

Health Level 7 standards

HL7 is a standard messaging format for data exchange between information systems. HL7 is commonly used to send orders from an order entry system to the LIS and to send LIS results to a clinical information system. It also is used for ADT systems to transmit relevant demographic

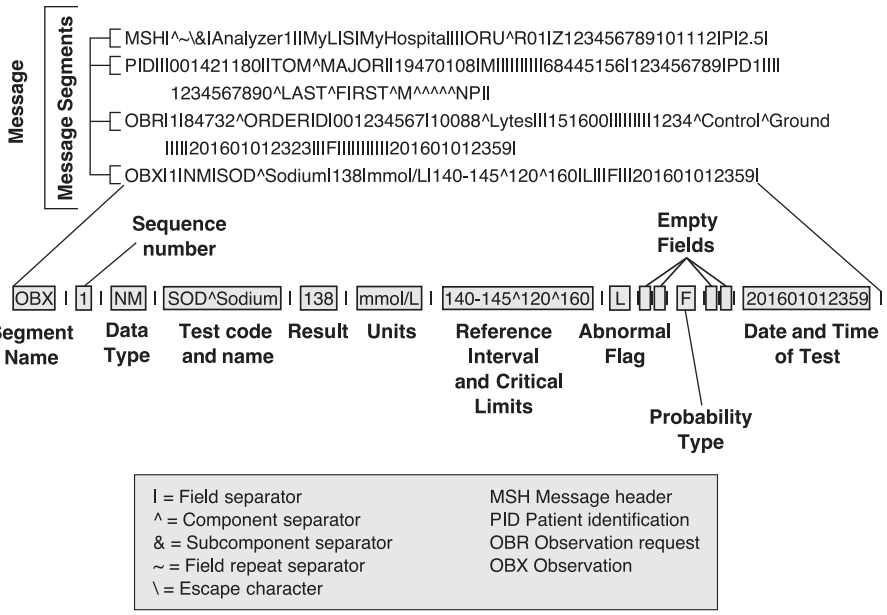


FIGURE 18.8 A sample Health Level 7 message. The message is a result message sent for a sodium result, with a value of 138, which is below the reference interval.

and patient status information (e.g., ADT-A01 is the code for admitting a patient). HL7 defines the types of messages that can be sent as well as the data that can be exchanged in each message. The use of extensible markup language (XML) for the encoding of data for storage, display, and transmission has become standard in many industries. Recently, healthcare has begun to utilize XML to supply the framework for clinical data, and XML-compatibility has been incorporated into newly developed HL7 standards.

HL7 message types include messages for test orders, test results, and ADT notifications. In addition, HL7 defines how the messages are to be exchanged and provides error communication guidelines. Each HL7 message is divided into segments [e.g., message header (MSH) segment and result/observation (OBX) segment] and the segments are further divided into fields (separated by the vertical bar character, "|") that contain the data to be exchanged (Fig. 18.8). For example, in the OBX (result) segment of the message, the OBX-3 field contains the test name and the OBX-5 field contains the test result value. In the example, the test name field in OBX-3 is divided by the caret character ("^") into two data elements, the test code, and the test name (e.g., 4498-2^Complement C4).

Logical observation identifier names and codes

Logical observation identifier names and codes (LOINC; see Table 18.4) is a standard for naming laboratory tests. LOINC attempts to provide a universal code to identify over 30,000 different laboratory-related items. LOINC specifies each test with a multipart name that includes the analyte name, property measured, timing (e.g., random time point or part of a timed collection), specimen type, and method (Table 18.3). LOINC was designed to enable disparate systems to more easily exchange and aggregate laboratory data. The utility of LOINC may include enhanced communication in integrated healthcare delivery networks and support for electronic health records, improved infectious disease surveillance, and improved laboratory test billing. The difficulty in utilizing LOINC is that there are many similar codes, and assignment of a specific code to a local laboratory test method can be challenging. LOINC is compatible with usage within HL7 messages and may be included in the result segment of an HL7 message.

Management of nonlaboratory information systems data

Unfortunately, no LIS capture all the data required to understand all aspects of the total testing process for a sample. Data that are typically not included and or maintained within the LIS include operational procedures, administrative data, environmental monitoring data, and spreadsheets. This section will discuss systems, requirements, and approaches for managing data that are not housed in the LIS. Non-LIS data can take many forms and are managed in a variety of ways; this section will focus on document control systems, quality and inventory control software, and networked file servers Topics are addressed in descending order of robustness, beginning with commercial off-the-shelf dedicated systems down to storage on shared network drives.

Standalone software systems

Inventory control, sample archiving, QC, and document control systems are all development tools that often have limited integration with the LIS. These systems are each complex independent software tools with many features and requirements. What follows is a discussion of the minimal requirements to link information housed in these systems back to patient test results.

Document control

Standard operating procedures are often housed in a standalone electronic document control system. One key requirement is that relevant standard operating procedures in use at the time of testing can be reliably retrieved and

TABLE 18.4 Logical observation identifier names and codes examples.

Code	Component (analyte) property measured	Timing	Specimen type	Scale
5802-4 Nitrite ACnc	ACnc	Pt	Urine	Ord
49563-0 Troponin I.cardiac	MCnc	Pt	Ser/Plas	Qn
2823-3 Potassium	SCnc	Pt	Ser/Plas	Qn

ACnc, Arbitrary concentration; *MCnc*, mass concentration; *Ord*, ordinal scale; *Pt*, random time point; *Qn*, quantitative; *SCnc*, serum concentration; *Ser/Plas*, serum or plasma.
http://s.details.loinc.org/LOINC/49563-0.html?sections = Simple.

reviewed. This is accomplished via a revision control system in which timestamped changes are linked to a user ID. Thus, at any time, a test result can be linked back to the appropriate version of a relevant procedure. The traceability between a result and the relevant procedures is vital for understanding the preanalytical, analytical, and reporting contexts of historical laboratory results.

Quality control

While some LIS programs have built-in QC modules, many laboratories opt to use a standalone QC system. Standalone systems typically have some advantages over a built-in program, including peer data availability, superior graphing capabilities, options, and customization, and associated modules for QC design, risk evaluation, QC rule selection, and uncertainty calculations. Built-in LIS QC programs are easier to interface and often have no added cost. However, they usually have limited graphing options and only basic functionality.

In general, QC systems record information about the QC material in use, such as name, catalog number, concentration, and lot number. The software allows for visualization of QC data over time. Equally important is automation of QC multirules and recording of timestamped metadata about the assay method such as maintenance, calibration, and reagent lot changes. This timestamped information is essential for assessing assay performance at the time of patient testing. These data may be retrospectively retrieved during laboratory investigations or audits. This traceability is required by regulatory bodies and legislation (e.g., Clinical Laboratory Improvement Amendments), which mandate that QC information for laboratory testing be directly linkable to the laboratory test result, and the LIS typically provides this functionality.

Inventory control

Commercial inventory control and sample archiving systems are not ubiquitous in clinical laboratory settings. Inventory control systems track information such as inventory location, lot, source, quantity, and expiration date. Inventory management systems ideally track expiry dates and flag inventory when quantities are below a set limit; however, many labs still use paper logs or spreadsheets to track inventory. Similarly, many labs have home brewed methods for archiving patient samples. Commercial sample archiving software tracks sample location, along with sample type, and patient identifiers.

Due to the maturity and sophistication of inventory control, sample archiving, QC, and document control systems, they appear amenable to integration with a modern LIS. In some cases, these systems may be included as a module within the LIS. However, a best of breed approach is possible through application interfaces between the LIS and the external systems; middleware programs are an example of this approach at work.

Servers, directories, and security

In the clinical laboratory, data that are not stored in the LIS or a commercial application are often stored on a networked drive with shared access. These networked drives reside on servers administered by the laboratory, department, or institution. Servers are computers that run an operating system configured for client/server networking. Client/server networking allows a program on a user's computer, called a client, to connect via a network to another computer called a server, through requests sent to the server's Internet protocol (IP) address. Servers are configured to provide access to shared resources, such as databases, and directories through open ports. These ports act as doors to access resources at the server's IP address. The server's operating system uses a network protocol to communicate with clients and fulfill requests. Server message block and network file system are two network protocols commonly used to allow clients to access files over a network. Storage of laboratory files on a shared network drive allows multiple users to access resources and facilitates scheduled backup of all stored data.

A networked file server allows user-controlled access to the servers file system. A file system controls how data are stored and retrieved by separating data into pieces and assigning each piece a name. Each named group of data is a file, and the structure and logic used to manage the groups of information are the file system. While not important for the file system, a consistent file naming convention is vital for users of the file system. Directory structures provide additional organization of files for the benefit of users and are also used to control user access to sensitive information as discussed in the next section.

Security and access control

Privacy of healthcare information is of utmost importance to the clinical laboratory. It is a good practice, regulatory requirement, and typically legislated requirement. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) stipulates how personally identifiable information maintained by the healthcare and healthcare insurance industries should be protected from fraud and theft. The General Data Protection Regulation (GDPR) sets a similar direction in Europe. Modern file systems provide two levels of security to assist laboratories in complying with HIPAA and GDPR: access control and encryption.

File and directory access are managed via access control lists. An access control list is a table of entries that specify user or group rights to a specific file or directory.

Users are identified by logging on to the client system with a user name and password. User and group rights can include a combination of reading, writing, executing, and deleting a file.

Encryption is valuable in cases where the access control has been bypassed. Encryption encodes information so that only authorized parties can access the unencrypted content. Encryption does not prevent interference such as deletion of information, but denies the content to unauthorized users. There are two main forms of encryption for networked directories: file system-level encryption and full disk encryption. File system-level encryption is a form of disk encryption, where individual files or directories are encrypted by the file system. In file system-level encryption, metadata, such as directory structure, file names, sizes, and modification timestamps are not protected.

In contrast, full disk encryption encrypts the entire partition where the file system resides hiding the metadata. File system-level encryption allows files to be accessed and decrypted based on access-control lists. It should be noted that storage of laboratory data on a networked shared drive also provides physical security in the form of limited access to the secure server room.

File types

There are two generic types of computer files: text files and binary files. Text files are the most simple and flexible file type and are commonly used for long-term storage in flat files. Flat files are simple file types that have no internal structure or hierarchy. Typical examples used in the laboratory include comma separated values (CSV) files and plain text files. The characters in a text file are encoded either using American Standard Code for Information Interchange (ASCII) or the more modern Unicode (Universal Coded Character Set) Transformation Format-8-bit (UTF-8). Unicode is a computing industry standard for the consistent encoding, representation, and handling of text expressed in most of the world's writing systems. UTF-8 is an 8-bit character encoding system that is compatible with the older ASCII standard. Text files allow only plain text content with very little formatting (e.g., no bold or italic types) and can be viewed and edited in simple text editors. The advantages of text files for data storage include universal adoption, cross-platform compatibility, human readability, and resistance to file corruption. The CSV format is an example of a text file commonly used for data storage in clinical laboratories.

A binary file is a file type that contains more than plain text and is usually stored as a sequence of bytes intended to be interpreted as something other than plain text. A Microsoft excel spreadsheet file (XLS) is an example of a binary file. An XLS file contains formatting information, formulas, and graphics in addition to plain text. Converting an XLS file to CSV strips away all of the binary data resulting in a plain text file. Binary file types are essential for the day-to-day running of a clinical laboratory, but are not well suited to long-term data storage. One of the main vulnerabilities of binary files for long-term data storage is the need for proprietary software for interpretation. Consequently, the laboratory will have to keep copies of software on hand in order to access the contents of binary files. This can be problematic as vendor support for operating systems and software changes and expires.

Secondary data use

When considering the uses and management of electronic laboratory data, it is important to separate primary from secondary uses. Primary uses include preanalytic workflow of entering patient demographics and lab test orders, as well as postanalytic workflow of entering and verifying results and generating invoices. In other words, primary uses are transactions directly in support of the performance of individual tests. Most of the functions of LIS and other clinical information systems would be considered primary. Secondary uses, on the other hand, include a wide range of activities that occur after the generation of individual test results (see Table 18.5) to address clinical, operational, research, and business questions. Most

TABLE 18.5 Examples of secondary uses of laboratory data.

Clinical uses	Generation of antibiograms
	Detection of outbreaks
Operational uses	Calculation of laboratory throughput
	Analysis of quality control trends
Research uses	Biomarker association studies
	Analysis of historic disease trends
Business uses	Calculation of cost to serve different customers
	Analysis to support marketing efforts

secondary uses involve aggregation and analysis of data across multiple patients and/or tests.

Data warehousing

In order to support fully the secondary uses of data, healthcare organizations often build data warehouses that are separate from LIS and other transactional systems. Data warehouses that are broadly inclusive of data from the systems across the organization are referred to as enterprise data warehouses. Single-department warehouses are typically referred to as data marts. Although most transactional information systems (i.e., those in support of primary data use) allow longitudinal queries and in theory could support a number of secondary uses, there are a number of technical benefits to setting up independent data warehouses. One is performance. Within a busy laboratory, every second counts, and you do not want technologist workflow to be impeded by an LIS that is processing a massive longitudinal query. LIS database architectures are optimized for rapid individual transactions rather than large longitudinal queries. Another issue is that for many types of secondary uses, it is necessary to merge data across multiple sources. For example, a research query might involve data from both the laboratory and the radiology department, and a business query might involve data from both the laboratory and a customer relationship management system.

Creating a data warehouse involves much more than simply replicating data from multiple systems into a single large database. A number of issues must be considered in order to make the database useful and reliable for its intended purposes.

- Consistent data definitions: Users of the data warehouse need to know the precise definitions of the fields in the database. For example, a data warehouse might include a field named "Specimen collection time" to record the time of phlebotomy. This field could be useful for calculating TAT for different tests. Suppose, though, that the laboratory might not have a process for recording the time of nurse collections and that, for these specimens, the "Specimen collection time" field contains the time at which the specimen is received in the lab. Alternatively, the data from nurse collections might be entering the data warehouse from a separate nursing information system in which the collection times are defined differently from those in the LIS. A database analyst wishing to use this data warehouse field to produce TAT reports would need to know these details in order to avoid producing misleading results. A similar and very common issue is that the use of a field in an upstream transactional system can change over time due to changes in the associated work processes. Ideally, during the construction of the data warehouse, a different field is created for each different version of the collection time in order to retain integrity of data definition. But at a minimum, all these data definition details need to be documented in a central location.
- Use of clinical data standards (see also section on data standards for the laboratory): Established medical coding and terminology systems such as LOINC, ICD-10, and Systematized Nomenclature of Medicine can be used to increase both the quality and the usefulness of a clinical data warehouse. When disparate source systems map their data onto common standards, it facilitates the integration of these sources into the warehouse. Even more important, standards enable analytic and visualization tools to produce more useful output with far less end-user effort.
- Transformation and cleansing: Data warehouse loading processes are often referred to as extract, transform, and load (ETL). An extract from a transactional system might have missing and/or nonsensical data due to errors in the originating systems or the extraction process. A data warehouse ETL process should, therefore, check for these conditions. For example, the ETL software could check birthdate fields to ensure that all birthdates are in the past, but no more than 120 years in the past. Transformation also includes normalizing data, e.g., converting fields, such as dates, to a common format.
- Timeliness: Most data warehouses are not populated in real time, but rather rely on daily, weekly, or monthly extracts from the source systems. This is not a problem for most research applications, but it can limit the types of clinical, operational, and business functions that can be supported by the data warehouse. To the extent possible, it is helpful to identify requirements for near real-time data prior to the construction of the warehouse and ETL processes.

Decreasing storage costs, together with newer and more flexible database management system, have created opportunities to create and use much larger and more complex data sets than those in the past. The term "Big data" is often used to refer to databases that have one or more of the "three Vs": Volume, Variety (particularly of nonnumeric data such as free text), and Velocity (rate of data entering the database). A fourth V, Veracity, refers to data that are difficult to cleanse (also referred to as noisy). Examples of big data in a medical laboratory setting may include data generated from microbiome sequencing, whole exome sequencing, data from continuous biosensors, and collections of textual physician notes.

Data analysis

The last decade has seen a proliferation of tools that support analysis and visualization of data within warehouses. These are often referred to as business intelligence (BI) software. This has had a democratizing effect, opening up analysis of larger and larger data sets to end users who do not come from IT backgrounds. Just as with statistical software, though, BI software in inexperienced hands carries a risk of drawing inappropriate conclusions due to misapplication of analytic techniques. At the other end of the spectrum is a growing demand for expertise in analysis of larger and more complex data sets. Data science is a growing job category that combines statistics and software engineering.

Machine learning and artificial intelligence

An important subset of data science is machine learning (ML), also referred to as artificial intelligence (AI). While this field has existed for many decades, it has grown rapidly in recent years due to increased availability of big data. A detailed discussion of ML is beyond the scope of this chapter. Essentially, it represents a wide range of algorithms such as neural networks, association rules, and decision trees/random forests, all of which are designed to identify subtle patterns within large data sets. In contrast to more traditional statistical techniques such as regression, ML can tease out a much wider range of (especially nonlinear) patterns in data sets. On the other hand, this power comes at the cost of the ability to explain the roots of those patterns. Furthermore, ML algorithms may identify patterns that reflect artifacts in the data set that do not exist in the underlying real world. It is thus critical that ML algorithms intended for medical use are subjected to extensive clinical validation in real-world settings before they are used for patient care (see Box 18.1).

Data privacy

The growing volume, extent, and interconnectedness of online patient data have created new challenges for privacy and security (see section on "Servers, directories, and security"). HIPAA, the primary regulatory framework for these issues within the United States, was written in the 1990s, an era with much less computing capacity and online data. This limitation is illustrated by HIPAA's provisions for deidentified data sets, in which key identifying fields are removed. Such deidentified data sets are generally considered exempt from HIPAA and can be shared with other parties, used for research, and used for commercial purposes. A significant problem is that under some circumstances, it can be possible to reidentify data within such a data set using external sources. To illustrate, consider a whole genome sequence that identifies a rare serious illness such as Huntington Disease, information that the patient may want to keep confidential. Under current common practice, a laboratory might "deidentify" this record by removing fields such as the patient name and medical record number, and converting the birth data to a birth year. At this level of deidentification, a third party such as an insurance company might still be able to cross-reference the data against other sources, such as billing records, to discover the identity of the patient with the disease.

Opportunities

Research

Laboratory data can be a powerful research tool, but like all tools must be used properly to achieve useful results. In addition to privacy and security issues (see above) that require institutional review board oversight, there is a major issue of controlling for bias when using uncontrolled, real-world retrospective clinical data. Real-world data are not necessarily either better or worse than data

BOX 18.1 The problem with AI black boxes in healthcare

In the late 1990s a number of ML models were built at the University of Pittsburgh and Carnegie Mellon University as part of a project attempting to predict pneumonia mortality based on inpatient data. These models identified a surprising pattern: pneumonia patients with comorbid asthma were predicted to have a lower mortality than other patients. This contradicted previous research that showed that asthma in the setting of pneumonia is an independent risk factor for death. It turned out that the data set for the ML models had come from electronic patient records at the University of Pittsburgh Medical Center, where pneumonia patients with a history of asthma were routinely admitted directly to the intensive care unit. This practice, while certainly beneficial to those patients, created a bias in the data set. If this bias had not been identified, and if the algorithm had been used to identify patients who could be safely discharged to home, then asthmatic patients could have died as a result. Fortunately, for patient safety, the algorithm was subjected to considerable study and validation rather than simply being trusted at face value.

from randomized controlled experiments, but they has their own types of biases that must be respected. Most of these issues fall into the category of correlation that does not prove causation. In the pneumonia mortality example (Box 18.1), asthma was correlated with lower mortality within the data set, in spite of the fact that asthma actually increases pneumonia mortality in the real world. The causative driver in this case, namely different ICU admission criteria for asthmatic versus nonasthmatic patients, was not captured in the data set.

On the other hand, well-designed clinical research based on retrospective data sets has a number of practical and ethical advantages over randomized controlled trials (RCTs). It is simply not feasible to subject all clinical questions to large RCTs, nor would most patients want all of their care to take place within RCTs. The most practical way to generate new clinical knowledge out of real-world settings, then, is through careful retrospective analysis. RCTs can then be used in complementary roles, to confirm findings and tease out sources of bias.

Summary

An understanding of the laboratory information management is necessary for the effective practice of laboratory medicine. Information systems, such as the LIS, have a direct impact on nearly every employee in the laboratory and provide the framework for laboratory workflow. In effect, the LIS is the engine for providing information to clinical decision makers. Standardization of interfaces and terminologies has led to the integration of the LIS into the healthcare information network. Beyond the core use of laboratory information for direct patient care, there are secondary uses. These include operations, research, and quality. Maximizing these uses requires effective storage and retrieval capabilities while maintaining security and privacy.

Acknowledgments

The authors gratefully acknowledge the contributions of Anand S. Dighe, MD, PhD, to the previous chapter (3rd edition).

Suggested reading

[1] M.J. Becich, J.R. Gilbertson, D. Gupta, A. Patel, D.M. Grzybicki, S.S. Raab, Pathology and patient safety: the critical role of pathology informatics in error reduction and quality initiatives, Clin. Lab. Med. 24 (2004) 913–943.

[2] G.F. Cooper, C.F. Aliferis, R. Ambrosino, J. Aronis, B.G. Buchanan, R. Caruana, et al., An evaluation of machine-learning methods for predicting pneumonia mortality, Artif. Intell. Med. 9 (1997) 107–138.

[3] D.F. Cowan, Informatics for the Clinical Laboratory: A Practical Guide for the Pathologist, Springer, New York, NY, 2005.

[4] Health Insurance Portability and Accountability Act of 1996 (HIPAA). <https://aspe.hhs.gov/report/health-insurance-portability-and-accountability-act-1996>. (accessed 19.08.19).

[5] <http://loinc.org> (accessed 01.01.19).

[6] <http://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf> (accessed 06.03.20).

[7] K.L. Imborek, N.L. Nisly, M.J. Hesseltine, J. Grienke, T.A. Zikmund, N.R. Dreyer, et al., Preferred names, preferred pronouns, and gender identity in the electronic medical record and laboratory information system: is pathology ready? J. Pathol. Inf. 8 (2017).

[8] Laboratory Analytical Workflow. <http://wiki.ihe.net/index.php?title = Laboratory_Analytical_Workflow#Background> (accessed 01-08-16).

[9] C.J. McDonald, S.M. Huff, J.G. Suico, G. Hill, D. Leavelle, R. Aller, et al., LOINC, a universal standard for identifying laboratory observations: a 5-year update, Clin. Chem. 49 (2003) 624–633.

[10] L. Pantanowitz, W.H. Henricks, B.A. Beckwith, Medical laboratory informatics, Clin. Lab. Med. 27 (2007) 823–843.

[11] L. Pantanowitz, J.M. Tuthill, U.G.J. Balis, Pathology Informatics: Theory and Practice, ASCP Press, Chicago, IL, 2012.

[12] M. Plebani, Exploring the iceberg of errors in laboratory medicine, Clin. Chim. Acta. 404 (2009) 16–23.

[13] K. Rodríguez-Capote, C.M. Balion, S.A. Hill, R. Cleve, L. Yang, A. El Sharif, Utility of urine myoglobin for the prediction of acute renal failure in patients with suspected rhabdomyolysis: a systematic review, Clin. Chem. 55 (12) (2009) 2190–2197.

[14] H. Wickham, Tidy data, J. Stat. Softw. 59 (10) (2014) 1–23.

[15] C.D. Hawker, Bar codes may have poorer error rates than commonly believed, Clin. Chem. 56 (10) (2010) 1513–1514.

[16] M.L. Snyder, A. Carter, K. Jenkins, C.R. Fantz, Patient misidentifications caused by errors in standard bar code technology, Clin. Chem. 56 (10) (2010) 1554–1560.

Self-assessment questions

1. An LIS dictionary is used for which of the following? (more than one may be correct)
   a. Translate information from the instrument to the LIS
   b. Standardize the drivers between the LIS and the EMR
   c. For looking up patient results in the clinical viewer
   d. Populate worksheets
   e. Provide the source information for dropdown lists
2. The LIS may be used for QA purposed by which of the following mechanisms? (more than one may be correct)
   a. Implementing autoverification rules
   b. Storing QC procedures
   c. Ordering test panels
   d. Generating patient reports
   e. Recoding MDIs
3. Which of the following are true regarding LIS connectivity?
   a. LIS transmits results to the CIS
   b. ADT transmits results to the LIS
   c. EMR transmits demographics to the LIS
   d. Instrument transmits results to the EMR
   e. HL7 receives results from the electronic health record
4. Which of the following is the correct sequence of events in an LIS?
   a. Test order -> Sample receipt -> Result from instrument -> Transmission to EMR
   b. Result Order -> Sample collection -> Result transmission -> Reporting
   c. Sample collection -> Result from instrument -> Result label -> Transmission to EMR
   d. Test order -> RFID -> Result receipt -> Transmission to interface
   e. Sample collection -> Batch testing -> Sample receipt -> Result Label -> Transmission to EMR
5. The role of laboratorians in the age of integrated digital health information is which of the following? (more than one may be correct)
   a. Provide test results
   b. Provide information
   c. Analyze and interpret information
   d. Provide connectivity between instrumentation
   e. Order laboratory tests
6. The CID is used in all of the following processes EXCEPT _____.
   a. the CID distinctly identifies every tube or container in a given specimen
   b. the CID is the key identifier that links order and result data in the analyzer and the LIS
   c. the CID encodes sample type information
   d. the use of unique CIDs is a requirement for the use of robotic automation systems
   e. unique CIDs are required by robotic automation systems
7. Which segment of an HL7 message contains laboratory testing information.
   a. EVN: event type
   b. DRG: diagnosis-related group
   c. OBX: observation/result
   d. ROL: role
   e. MSH: message header
8. LIS data are amenable to what data manipulations techniques:
   a. filter, transfer, congregate, and sort
   b. sort, filter, aggregate, and transform
   c. collect, sort, file, and store
   d. sift, shuffle, pile, and dump
   e. ignore
9. Which of the following is NOT true of instrument interfaces?
   a. Most instrument−LIS interfaces are based on standards and protocols defined by the ASTM.
   b. To interface an instrument, the LIS must be configured with a device driver.
   c. Unidirectional interfaces solely send orders from the LIS to the analyzer.
   d. Bidirectional interfaces permit two-way exchange of data between the LIS and the analyzer
   e. Bidirectional interfaces may send orders/demographics to the instruments in batches or dynamically.
10. Which of the following is NOT a property of well-designed electronic result reports?
    a. Recent results are highlighted
    b. Related tests are grouped together
    c. Results can be filtered
    d. Interpretive comments are hidden from view
    e. Results can be trended graphically
11. Which of the following is not a function of a laboratory information system?
    a. Relational database
    b. Data exchange interface
    c. Laboratory process automation
    d. Data warehouse
12. Which of the following is not true for full disk encryption?
    a. Decrypts content based on access control lists.
    b. Encrypts file metadata.
    c. Encodes information so that only authorized parties can access the unencrypted content.
    d. Prevents deletion of encrypted information.

13. Which of the following best illustrates a secondary use of laboratory data?
- **a.** Research study using the past 4 years' worth of data to look for correlations among common biomarkers.
- **b.** Retrieval of a patient's lab results in response to a physician's query.
- **c.** Submission of lab results from an instrument to the laboratory information system via an interface.
- **d.** Review of results from a run of an immunoassay prior to verification.

14. Which of the following statements best represents an advantage of retrospective data analysis for clinical research versus prospective data collection?
- **a.** Retrospective data allow for greater control in selecting patients for a study.
- **b.** Retrospective data allow for greater blinding of participants with regard to therapy selection.
- **c.** Retrospective data are preferred by regulatory agencies such as the FDA.
- **d.** Retrospective data can be aggregated at less expense than conducting a prospective trial (correct answer).

Answers

1. c–e
2. a
3. a
4. a
5. b and c
6. c
7. c
8. b
9. c
10. d
11. d
12. d
13. a
14. d

Chapter 19

Point-of-care testing

James H. Nichols
Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, United States

Learning objectives

After reviewing this chapter, the reader should be able to:

- Define point-of-care testing (POCT).
- Identify the essentials of quality management for reliable POCT results.
- Describe POCT limitations.
- List the resources available to assist staff in managing their POCT program.

Introduction

Point-of-care testing (POCT) is laboratory testing conducted close to the site of patient care. There are a number of portable POCT devices on the market, and the available menu of analytes is extensive, including glucose, urinalysis, pregnancy, occult blood, electrolytes, blood gases, creatinine, urea, microalbumin, hemoglobin A1c, drugs of abuse, therapeutic drug monitoring (lithium), as well as testing for infectious diseases like streptococcus, mononucleosis, influenza, and HIV. Even coagulation testing [prothrombin/international normalized ratio (INR)], enzymes [alanine aminotransferase (ALT)/aspartate aminotransferase (AST)], and cardiac markers [B-type natriuretic peptide (BNP)] are available on the Clinical Laboratory Improvement Amendments of 1988 (CLIA) waived devices. CLIA waiver is granted to laboratory tests that the US Food and Drug Administration (FDA) and Centers for Disease Control and Prevention (CDC) have determined to be so simple that there is little risk of an erroneous result. CLIA waived tests have minimum regulatory requirements, so a majority of POCT is CLIA waived, allowing the test to be performed in a variety of settings by a diverse array of staff. However, as with any test, POCT, even CLIA waived testing, has limitations and poses risk of incorrect results and patient mismanagement if incorrectly performed or poorly managed. POCT is also sometimes referred to as near-patient testing, bedside testing, ancillary testing, and satellite testing, because the POCT devices can be used in many different patient-care settings: hospital acute and critical care units, operating rooms, emergency departments, physician's offices, visiting nurses, patient home self-testing, ambulances, and cruise ships. POCT devices have even traveled on the space shuttle.

POCT has a number of advantages (Table 19.1) compared with central laboratory testing. POCT devices use a small volume of sample, and the sample does not need to be processed. POCT is particularly useful in neonates, small babies, and those with increased risk of blood loss from phlebotomy. POCT devices can analyze whole blood, urine, and other body fluids. Test results are rapidly available while the clinician is examining the patient. There is no need to transport a specimen to a laboratory and wait for results to come back. POCT thus works with the clinical flow of patient care, because clinicians can order a test, receive results, and institute clinical action in the same visit. This provides the opportunity for faster therapeutic intervention and potential for improved patient outcomes.

However, POCT has some disadvantages (Table 19.1) as well. POCT is more expensive than traditional laboratory testing, because POCT reagents are single-use tests, while central laboratory reagents are packaged in bulk for high-volume analysis. POCT in hospital settings is often not performed by lab personnel, resulting in challenges in identifying deviations in preanalytic, analytic, or postanalytic components of testing. Documentation is an additional concern, because POCT may be performed and results not interfaced or documented in the patient's electronic health record. Clinical staff may be unfamiliar with the many laws and compliance requirements that apply to laboratory operations both within the United States and internationally.

Quality point-of-care testing

Quality POCT is the delivery of results that are safe, reliable, and trustworthy for patient care. Quality implies the appropriate use of technology to meet clinical needs. This

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00019-3

TABLE 19.1 Point-of-care testing advantages and disadvantages.

Advantages
• Immediate results—no lab transportation
• Small blood volume
• Wide menu of tests available
• Whole blood, urine, and other body fluids analyzed without processing
• Works within clinical patient flow
Disadvantages
• More expensive than traditional laboratory tests
• Quality is questionable as anyone can run the analysis
• Difficulties with regulatory compliance, billing, and documentation
• Patient-focused staff with little formal education/experience in laboratory testing

TABLE 19.2 Example point-of-care testing program for an academic medical center.

Method	Sites	Devices	Operators
Glucose meters	53	213	4500
Hemoglobin A1c	10	21	50
Fecal occult blood	38		1400
Gastric occult blood	6		500
Urine dipsticks	43	28	650
Urine pregnancy	41		600
Serum pregnancy	1		25
Rapid strep	20		300
Mononucleosis	1		3
Influenza	14		225
pH paper	2		50
Blood gas portable	10	15	400
Hemoglobin	3	4	50
Avoximeter	2	7	30
Prothrombin time/INR	7	7	60
Activated clotting time	7	30	150
Heparin management	1	3	6
Basic metabolic panel	6	9	275

involves choosing the right method, POCT device or central laboratory analysis, to achieve the best outcome for a particular patient at a specific point in their pathway of care. This might involve selecting between a glucose meter, glucose on a blood gas analyzer, or glucose in a central laboratory, depending on the clinical situation. In addition, test results are legal documents and all of the quality control (QC), training, instrument validation, troubleshooting, and other quality assurance practices and documentation ultimately provide defensible evidence of regulatory compliance should a test result ever come into question. Above all, the reputation of the laboratory depends on the quality of its product, the test result.

Unfortunately, there are a number of quality concerns with POCT, which have been raised in the scientific literature. Complaints about glucose meters for patient self-testing, for instance, represent the largest number of complaints filed with the FDA for any medical device, with over 3200 incidents including 16 deaths [1]. POCT devices can be reservoirs for nosocomial and antibiotic-resistant organisms in the hospital, and infections have been transmitted by poorly cleaned urinometers and blood gas devices when carried between patient rooms [2]. Patients at nursing facilities in Mississippi, North Carolina, and California were diagnosed with hepatitis B infection transmitted in association with blood glucose monitoring from shared spring-loaded lancets, glucose meters, insulin vials, and poor hygiene between the patients [3]. In a two-state pilot study conducted by the Department of Health and Human Services (DHHS) on data from state laboratory inspectors, 77% of the 175,000 CLIA enrolled laboratories had no direct oversight by someone with laboratory training [4]. Most of these laboratories were physician office practices, and 50% of laboratories in Colorado and Ohio did not follow manufacturer's instructions nor did they have package inserts available to staff performing the testing. Other laboratories were cutting occult blood and urine dipsticks into segments to save costs and extend their use. In an expanded study conducted in eight states, 64% of laboratories (173/270) surveyed did not have manufacturer's instructions or failed to follow the instructions [5]. Twenty percent of laboratories ($n = 53$) were not performing QC as instructed, and 19% of laboratories ($n = 51$) did not provide training to staff or evaluate the staff's ability to perform testing adequately.

POCT is a complex system. In a central laboratory, there is one location and testing is limited to a few analyzers. Staff is also limited to a pool of medical technologists and medical laboratory scientists who have significant training and skills in laboratory analysis. Staff in a central laboratory has focused attention on testing. Their primary task is specimen analysis and producing quality test results, without the distraction of patient care responsibilities. POCT, on the other hand, is conducted at dozens of locations, with hundreds of devices and thousands of operators (Table 19.2). Getting all operators to

perform a test consistently, in the same way every time, is difficult. POCT is performed by clinical staff whose primary focus is patient care. Nurses and physicians are only secondarily focused on performing the necessary steps for QC, calibration, maintenance, cleaning, and quality assurance of POCT devices. There is the assumption that if a POCT device generates a result, then that result must be correct; otherwise, the device would not have generated a value. However, there are many factors that can affect the quality of the POCT results.

The DHHS studies from state inspectors uncovered quality concerns with the performance and staff training in a significant percentage of sites performing POCT, mostly physician offices [4,5]. Most physicians and facilities were eager to change once the issues were brought to their attention but had no resources or background to design the training and ongoing quality assurance program. Laboratory professional consultation was needed to improve the regulatory compliance of these sites and guarantee the quality of POCT results. For many hospitals and health systems, this consultation is available through a medical technologist who specializes in POCT quality assurance, called the "point-of-care coordinator." The point-of-care coordinator is the laboratory liaison to the nursing unit to assist with operator training, instrument validation, documentation, and quality improvement. For some health systems, the point-of-care coordinator is extended to physician office laboratories and home nursing networks under the health system umbrella, whereas other private physician offices hire pathologists, doctoral-level laboratory directors, or point-of-care coordinators as paid consultants to assist with their POCT regulatory compliance. The point-of-care coordinator is thus a specialized medical technologist who has the additional regulatory knowledge, communication, and management skills to help POCT sites improve the quality of their testing and meet regulatory guidelines.

Point-of-care testing regulations

The challenge of managing POCT is the complexity of ensuring regulatory compliance at dozens of sites, with hundreds of devices and thousands of operators. Each site must supervise equipment, personnel, and safety practices comparable with performing the testing at a single laboratory location; therefore POCT multiplies the many laboratory issues, like validation and monitoring of devices, reagent lots, expiration dates, and temperatures across dozens of sites. Inspectors use the phrase "site neutral" to describe the test requirements that are independent of where the test is conducted, and POCT sites are treated like independent laboratories regardless of location.

The quality of POCT, like other laboratory tests, is regulated by the federal CLIA law and inspected through centers for medicare and medicaid services (CMS) and the state departments of public health. CLIA sets the minimum requirements for laboratory testing based on the complexity of analysis, instrument maintenance, and result interpretation. Tests are categorized as waived, moderate, or high complexity with increasingly higher levels of requirements for operator education, training, method validation and quality documentation. The Joint Commission, the College of American Pathologists (CAP), and the Commission on Office Laboratory Accreditation are private accreditation agencies that are deemed under CLIA to inspect and accredit labs that meet the minimum CLIA standards in addition to their individual agency recommendations. Although each of the accreditation agencies varies in the details and stringency of their recommendations, all of the agencies require some level of equipment validation before use of the test on patients, written procedures, operator training and competency on the specific test, as well as ongoing management of patient results (with appropriate normal reference ranges or interpretive comments) and safe handling and disposal of biohazardous patient specimens. CLIA compliance is enforced by blocking Medicare and Medicaid payments for noncompliant laboratories and levying fines against the organization and laboratory director, while accreditation agencies can pull a laboratory's accreditation, institute fines, and publicize citations for laboratories with quality concerns. Despite the sale of many POCT devices in local grocery stores and pharmacies, a physician cannot simply walk into a store, purchase a device, and start testing patients without first ensuring licensure of the site performing testing and documentation of quality assurance for regulatory compliance.

POCT regulations cover laboratory testing conducted for patient care decisions and medical management. The portability and ease of use of POCT devices make them convenient for use in research protocols; however, research testing is exempt from the CLIA regulations, but only if no patient care decisions are made from the test results. Staff should be warned that the CLIA guidelines still apply to the use of a POCT device as part of a research protocol if a medical decision is made or patient care is changed based on the test result. For example, the use of a pregnancy test to determine whether a patient can enroll in a research study and potentially receive a study drug is a medical, patient care decision, so all of the CLIA and accreditation guidelines would apply to the pregnancy test. Alternatively, if the pregnancy test was performed as part of the study and none of the test results were communicated to anyone caring for the patient, no change in medical management could occur based on the test. This testing would be considered purely research. When results are recorded in a log to be reviewed after the study completion, the results cannot be utilized for

patient care and the test is classified as research; thus CLIA does not apply in this case. A good rule of thumb to determine if the regulations apply is to answer the question, "Will a change in treatment or management occur based on the result of this point-of-care test?" For the pregnancy test, a decision may be made to include the subject in the study if the test is negative and to exclude the subject if the test is positive. This constitutes medical management, in which a clinical decision (to enroll in the study) is made based on a test result. CLIA regulations therefore apply in this case, but do not apply when the test results are protected in such a way as to prevent the use of the result by any clinicians actively caring for the patient.

Documentation is fundamental to all of the accreditation requirements, and data records and logs are considered the supporting documentation for regulatory compliance. Some devices provide automated electronic capture of the necessary data (patient and operator identification, date, time, device serial number, reagent lot number, expiration dates, etc.) at the time of analysis. However, over 50% of POCT devices are manual, visually interpreted dipsticks and kits that require handwritten records or manual input into an electronic medical record. This has led to an increase in the utilization of automated readers for manual dipsticks and lateral flow tests that can automate the interpretation of results and be interfaced with the patient's electronic medical record. The amount of data requiring review from POCT can be overwhelming, and computerized POCT data management can assist personnel in regulatory compliance, billing, and review of data for quality improvement. Point-of-care coordinators can assist personnel with interpretation of the regulatory guidelines, development of a quality program, and preparation and review of the necessary documentation to meet the specific accreditation guidelines. Some laboratories may be accredited by multiple agencies and must prepare documentation to support inspection against multiple sets of recommendations. Federal agencies like CLIA may deem equivalence to another agency with stricter recommendations such as New York State, the Joint Commission, or CAP, for instance. Agencies are continuously updating their recommendations, so point-of-care coordinators and staff must keep current on the most recent changes that may affect their POCT operations.

Point-of-care testing connectivity and interfacing

Computerized data management automates the collection and review of the volumes of data acquired by POCT. Newer POCT devices have computerization that ensures regulatory compliance by collecting pertinent information at the time of testing and prompting the operator through the correct sequence of steps to perform the test consistently. Data stored within the device can be downloaded to a computer database or POCT data manager, by placing the device into a docking station that connects to the Internet, via modem, or wirelessly to a remote POCT data manager. The POCT data manager can also interface with a laboratory information system (LIS) and a clinical data repository (CDR), such as a hospital information system (HIS) or electronic patient medical record, and other clinical information system (CIS) for permanent storage of the test results.

The POCT data manager has multiple functions. The data manager communicates bidirectionally with devices. From the device, the data manager collects test information and associated date, time, patient identification (ID), operator ID, lot and serial number fields, as well as any associated result flags and comments. The data manager can also send current lists of competent operators and update any reagent and control lots, expiration dates, and other information to the device remotely. The data manager acquires control results to document successful performance of QC at the required frequency of testing, as well as any corrective actions for failed QC. POCT data managers use patient identification to search against active patients in the hospital or LIS to transfer patient results to those systems for permanent storage. Newer POCT devices have the capability of positive patient identification. These devices acquire admissions/discharge/transfer information on patients through the data manager and can display the patient's name on the POCT device before analysis. The operator then confirms the patient identification by entering a second patient identifier (like birth date) before the device will allow testing. Positive patient identification reduces the chance of ID entry errors that can prevent results from being transferred to the correct patient's medical record. The data manager thus acts as a central processor for information coming from POCT devices to the patient's medical record and for information required by a device for continued operation. POCT data managers automate the review of data and reduce the amount of labor required to manage POCT operations.

Historically, each manufacturer developed their data manager systems in isolation, meaning that different POCT devices required different physical cables to connect to the same data manager, and even language and communication protocols for data were not shared between different devices and manufacturers. This complicated and significantly increased the cost of implementing and changing devices because of the need to purchase additional computer systems and interfaces to accommodate the data transfer. The lack of common communication standards led to the development of the POCT1 standard for point-of-care connectivity by the

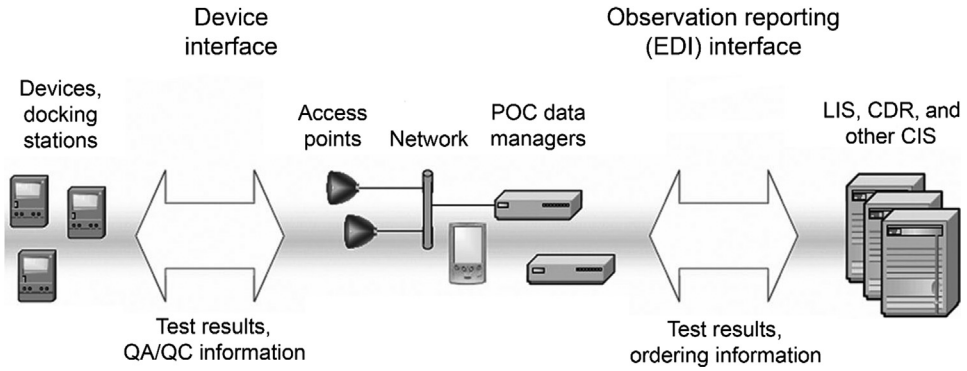


FIGURE 19.1 The POCT1 device connectivity standard describes two interfaces. The device interface allows bidirectional communication between the point-of-care testing devices and the point-of care data managers through docking station access points using the Internet, wireless, or serial communication. The observation reporting interface is an electronic data interface that transfers patient test results and order information between the point-of-care data managers and the laboratory information system, clinical data repository (an electronic medical record), or other clinical information system. *Reproduced with permission from Clinical Laboratory Standards Institute. Point-of-Care Connectivity: Approved Standard POCT1-A2. CLSI, Wayne, PA, 2006, p. 306.*

Clinical and Laboratory Standards Institute (CLSI) [6]. POCT1 was developed by a group of vendors, consumers, and government representatives to enable the seamless information exchange between POCT devices, electronic medical records, and LIS. This standard defines the physical connections as well as the communication protocols for data transfer between the POCT devices and the data managers (the device interface), as well as between data managers, and the LIS, CDR, or CIS through the electronic data interface (Fig. 19.1). The POCT1 standard is based on the Institute of Electrical and Electronics Engineers and Health Level Seven standards for medical information transfer, and representatives from these groups worked with CLSI in the development process. There are currently many devices that comply with the POCT1 standard, and several POCT data manager systems currently use this standard to connect to multiple POCT devices.

Point-of-care testing quality assurance programs

Compliance with regulatory guidelines requires an understanding of the regulations, good communication among all the staff involved in POCT, organization, and planning. The formation of a POCT committee, an interdisciplinary committee to oversee POCT approvals and quality concerns, can assist in the development of a quality assurance program and provide a discussion forum for issues as they arise. It is important to have representation from all groups involved in POCT. Administration, laboratory, nursing, and physicians are key individuals to have on the POCT committee, but representation from purchasing, distribution, pharmacy, and nutritional services may also be important when specific issues arise. POCT committees develop high-level policies that can cross departmental and institutional boundaries in hospitals and health systems with departmental or multihospital administration silos.

A key function of the POCT committee is to review and approve requests for new tests and to address compliance concerns with existing testing. Due to its interdisciplinary composition, the committee can balance opinions, depersonalize judgments, and deflect the political heat that may arise from decisions to remove testing or decline the implementation of new testing on specific nursing units or locations. The committee can take the blame for a decision rather than an individual like the POCT coordinator or laboratory director. Representation from purchasing and distribution can alert the committee to requests for tests outside of the committee approval process if staff try to go around the committee. Other ancillary staff, like nutritionists who may change diet based on glucose or electrolyte test results, may also be useful to have represented on the POCT committee.

Pharmacists are becoming more involved in POCT, not just because of drug dosing based on POCT results, but because pharmacies are now stocking POCT devices and pharmacists in some hospitals are starting to perform POCT as an alternative to nursing staff. Pharmacy-based clinics, in Britain and America, such as minute clinics and urgent care centers, are starting to offer POCT as part of a rapid care pathway designed to facilitate same day visits when patients cannot get a same-day appointment with their primary care physician. Patients are often seen by nurse practitioners, and POCT is offered to make a diagnosis (like rapid strep testing) and get treatment (such as an antibiotic prescription) in the same visit. POCT in these clinics is conducted at the request of the clinician or pharmacist, but in some locations, patients can even order their own tests and have them conducted by the

pharmacist (direct-access testing). Pharmacists, because of their training and experience with POCT, can provide a quality result that can be trusted more than patient self-testing at home, and pharmacy testing may be more convenient for patients than having to wait to schedule a physician visit, get a test order, travel to a phlebotomy station that will then transport a specimen to a central laboratory, and further wait for a test result to be communicated back to the physician where a call can be made to the patient to institute treatment. Think of the urine pregnancy test: if the patient performs the test at home, they must then go to a physician for repeat testing by a central laboratory to verify a positive result. Visiting a pharmacy creates a reliable test result that can be trusted for ongoing care, because the pharmacy must follow laboratory quality regulations when performing POCT compared with home self-testing, which has no quality requirements.

A POCT quality assurance program encompasses all of the device standardization, validation, training, competency, and operational aspects of performing POCT. POCT should be considered a part of patient care within the institution rather than an ancillary laboratory service, so POCT should be incorporated into the overall institutional quality improvement plan. Quantitative measures of POCT performance, such as compliance errors with monitoring temperatures, failure to recognize reagent expirations, or corrective actions after control failure, can be tracked and sorted by site to determine persistent problems and trends (Table 19.3). Nursing units and sites with recurrent quality issues should implement additional follow-up, closer supervision, or other actions to improve performance. Accreditation inspectors like to see trend graphs, demonstrating the identification of a problem, intervention, and improvement of performance over time. Accreditation agencies used to schedule periodic inspections with advance notification; however, the Joint Commission and CAP now utilize unannounced inspections. Staff must always be ready for an inspection and will no longer have time to prepare records and quality documentation in advance. Use of quantitative performance monitors and integration of the POCT program into the institution quality improvement plan are a good means of identifying and continuously improving on program weaknesses. Development of an individual quality control plan (IQCP) can identify risk of errors and define a quality assessment for tracking ongoing performance. Baseline performance assessment and monitoring trends of improvement over time also provide excellent documentation whenever a POCT result is questioned, regardless of who questions a result—a clinician, a patient, a staff member, or an outside inspector.

TABLE 19.3 List of possible quality improvement performance monitors for point-of-care testing.

Successful QC
• QC documentation • Number of errors where wrong QC analyzed (e.g., high control analyzed as low) • Percentage of QC that fail • QC outliers with comment • Failed QC with appropriate action (patients not tested)
Utilization (number of tests per site or device)
• Tests billed versus tests purchased • Single lots of test and QC in use at any time
Compliance
• Untrained operators performing testing • Clerical errors or data entry errors during testing or result reporting • Medical record has results reported with reference ranges • Expired reagents and QC/reagent bottles and kits dated appropriately • Refrigerator temperature monitored • Proficiency testing successful • Action plan in response to site compliance deficiencies

Interdisciplinary communication

The management of POCT requires the development of a quality assurance program and clear communication of that program to everyone involved. The objectives of improved patient care through POCT and the path to achieving this goal need to be clearly defined. Likewise, the fundamentals of regulatory compliance and quality improvement must also be spelled out. Unfortunately, clinicians perceive POCT from a different perspective than the laboratory. Technology and nursing has a kind of "love−hate" relationship. Taking the optimistic view, technology is intricately linked to the science of nursing and physician practice. POCT provides more rapid clinical decision-making, moving patients through the system (the hospital or physician's office) more quickly; improving the efficiency of delivering care; and decreasing patient wait times. However, from the pessimistic viewpoint, technology detracts from the art of nursing and does not integrate well into the daily patient care responsibilities. Compliance with regulatory demands and quality assurance protocols is time- and labor-intensive for clinical staff and takes the focus of nurses and physicians away from the patient. Physicians are also experiencing increased pressure to examine more patients while conducting successful research programs, maintaining resident education, and developing unique programs in their discipline for academic advancement. Nurses are encountering increasing pressure to cross-train and diversify, take on more roles, multitask, and increase the use of unlicensed personnel for some aspects of patient care.

Nurses and physicians must be responsible for the physical, emotional, and spiritual care of the patient. These caring and nurturing roles are in direct contrast to the scientific rigors of the laboratory that include accuracy, precision, and quality focus.

Successful POCT management requires an appreciation of different viewpoints, mutual respect, and compromise. There are always multiple ways to achieve a goal. Conflict with POCT frequently arises from the laboratory, dictating how clinicians should act in response to a problem without considering all perspectives. As compliance issues or other differences of opinion arise, participants need to look beyond the paradigms of their discipline and think outside the box. In any field, there are shared sets of assumptions or paradigms. Paradigms determine what is taught, the methods, and how findings are interpreted. Nursing has one set of paradigms, the laboratory has another, and physicians have a third set of paradigms. When a compliance issue is noted, like an unapproved POCT device, the laboratory may interpret this as a willful breach of policies and failure to go through the POCT committee approval process. However, the physicians may perceive this only as an attempt to improve the care of their patients and improve turnaround time of results. They may not be aware of the stringency of laboratory regulations and the need for committee approval or device validation before using a POCT device. Nursing, on the other hand, was conducting the test, because the physician ordered it and was only following instructions. The problem here is a lack of understanding about the regulations governing POCT and a failure to communicate effectively the issues. A clear discussion of the POCT committee policies and the reasoning behind the policies will be more successful than dictating compliance. When staff understand why they must follow policies, they will be more likely to comply with QC performance, refrigerator temperature monitoring, and seeking approval before purchasing new POCT devices.

Self-management

POCT is a decentralized process, and the laboratory cannot be at every site 24 hours a day to monitor test performance. The laboratory has to delegate responsibility for POCT to the clinical staff and trust that nursing and physicians will properly manage the testing process. In this model, the laboratory takes a central oversight role for POCT, but the clinical staff takes responsibility for the day-to-day operations such as purchasing reagents, discarding outdated tests, validating staff competencies, and performing QC and other functions needed to maintain testing on the medical unit. This laboratory−clinical partnership is based on mutual respect and shared responsibility for POCT quality.

Successful POCT is based on self-management. Clinical units make a decision that faster testing will improve patient care and must allocate the resources necessary to support the testing, while the laboratory provides all of the policies, procedures, and information required for the staff to succeed and continuously improve their performance. One means of distributing POCT information to many sites within an institution is through a dedicated website for POCT. A website is easily accessible from anywhere within a health system with only a computer, web browser, and internet connection. An institutional POCT website may contain contact information for the lab director, point-of-care coordinator, and support staff as well as training checklists, policies and procedures, compliance reports, minutes of the POCT committee, current POCT projects, and links to the FDA laboratory safety tips or current CAP/Joint Commission accreditation checklists. A website could also contain a list of POCT contacts on each of the clinical units and frequently asked questions written in a language that the clinical staff can understand and relate. The website could contain compliance trends by unit of their performance monitors, so units can compare their performance to others and help prioritize their action plans. Self-management encourages clinical staff to take charge of their own testing.

Many institutions have adopted electronic document control systems that can be an alternative to creating a dedicated POCT website. Electronic document control systems allow the procedures to be available on the nursing units from any of the clinical workstation computers, so staff does not have to hunt for a hardcopy of a procedure when conducting testing. Electronic document systems remind staff for required annual review of policies and provide an easy means of distributing policy revisions to all the sites performing POCT simultaneously. Most importantly, POCT policies and procedures can be integrated with other nursing policies in format and wording, emphasizing the role of POCT in patient care rather than as a "laboratory" service.

Institutions with a large number of staff have also adopted electronic training and competency systems. These programs can track operator competency and send messages to those pending expiration of their training. Training/competency systems can utilize slides, video with sound, exams, and other tools to refresh staff on key points to remember when performing POCT and ensure that these points were received and understood. CLIA and accreditation agencies require six elements of competency to be evaluated annually including:

1. Direct observation of routine test performance
2. Monitoring the recording and reporting of test results (including critical results)

3. Review of intermediate test results, worksheets, QC records, proficiency test results, and preventive maintenance records
4. Direct observation of preventive maintenance and function check performance
5. Assessment of test performance using previously analyzed specimens or proficiency testing samples
6. Assessment of problem-solving skills

Many of these elements can be assessed through an electronic training/competency system with an exam that includes troubleshooting questions, but direct observation of test performance and comparison of test results requires one-on-one time of a staff trainer with each POCT operator. Observing test performance cannot be automated through an electronic training system.

Staff members who perform testing are wholly responsible for their success as well as their failure. By positioning the laboratory as a resource for POCT and distancing the laboratory from the performance aspects of the POCT program, staff members tend to help each other, communicate more about their issues, and work more as a team than when the laboratory is mandating compliance and dictating practice. Each unit can find unique solutions to issues that work best within the unit's workflow. Issues are brought to staff attention by the laboratory and staff is encouraged to inspect their own practice and determine the solution that operationally works best for them. For instance, two levels of QC must be analyzed each day of testing, but this can be accomplished in a number of ways. Nursing units should rotate the responsibility among different staff, but a nursing unit might choose to have the first shift run QC, because it works better into their patient rounds, while another unit might find that night shift works best when staff are not as busy. By facilitating communication between the POCT sites and the encouraging staff to solve their own problems in a regulatory compliant manner, the laboratory becomes a resource of information to help staff improve their own program. Staff can then find the best ways to integrate POCT into their current workflow without the laboratory interfering. This improves interdisciplinary relations between the clinical sites and the laboratory and promotes collegiality rather than animosity.

There are many POCT resources that are available to assist staff with managing POCT. A journal, *Point of Care: The Journal of Near-Patient Testing & Technology* (www.poctjournal.com), is dedicated to basic research, method validation data, regulations, and other topics related to POCT management and device technology. *The Journal of Analytical Laboratory Medicine* also publishes peer-reviewed research and offers case studies, focused studies, and laboratory reflections involving POCT. Several textbooks are also devoted to POCT [7,8], and the American Association for Clinical Chemistry (AACC) has a community forum, the AACC Artery, to discuss POCT issues (www.aacc.org/community/forums). AACC also offers a POC Bootcamp and certification program for coordinators who are managing POCT programs. There are also a number of regional point-of-care coordinator groups throughout the country, and their meetings, activities, and original articles are available on a POCT website (www.pointofcare.net). Point-of-care coordinators and staff involved in POCT thus have a variety of resources and ways to contact others involved in POCT to get answers to their questions, seek advice, and get feedback from staff who are experiencing similar challenges.

Analytical performance

Although regulatory compliance and management is a significant aspect of POCT, the analytical performance of devices also contributes to the overall quality of results and clinical applicability of a test. Poor testing will result when staff fail to follow manufacturer's instructions on a device with excellent analytical performance; it is the same as if staff do everything right, but picks a device with poor analytical performance. Technical performance and management practices both contribute to quality POCT. Therefore institutions must be concerned about picking the right test for the specific patient as well as ensuring that the test is conducted properly, and QC and quality assurance practices are followed.

The required performance characteristics of a device must match the clinical need of the patient and the question being asked by the physician. POCT devices are different methods than core laboratory instrumentation. Each method has unique bias, imprecision, and analytical interferences that may limit its clinical utility. Glucose meters, for instance, have greater imprecision and are susceptible to different interferences than central laboratory methods. Despite recommendations by the American Diabetes Association for agreement of glucose meters within 5% of a central laboratory method, none of the currently marketed glucose meters can meet these performance standards [9]. Glucose meters are thus not recommended for use in screening or diagnosis of diabetes [10,11]. The variability of glucose meters limits their use to management of patients already diagnosed with diabetes. CLSI recommends that glucose meters agree with a comparative central laboratory method within ± 12 mg/dL for results <100 mg/dL and $\pm 12.5\%$ for results ≥ 100 mg/dL [12]. Glucose meters are too variable for use in analyzing specimens collected after a glucose tolerance test for diagnosis of diabetes, and meter variability also limits the utility of meters in health fair and shopping mall screening programs. Device performance requirements can change with

the concentration of analyte and the clinical application (screening, diagnosis, or management).

Operators are a key part of device performance, and there are several studies indicating poorer performance by staff without laboratory training than operators with training. The CDC compared performance on proficiency testing samples sent to laboratories performing POCT [13]. Proficiency samples are specimens of unknown concentration analyzed like patient specimens where the results are graded and compared with all other laboratories performing the same test. The CDC noted significantly better performance on proficiency samples for medical technologists working in accredited hospital laboratories than nonlaboratory staff performing testing in physician office laboratories. Intermediary performance was seen at those sites performing testing with laboratory-trained operators. Training improves performance, and the Oregon Health Sciences University noted significant improvement in both the performance of glucose testing and interpretation of results after a standardized training program [14]. Performance not only improved immediately after training but continued to improve consistently for the following several months that were monitored by the study. Operators and the training of operators can thus affect the quality of POCT results as well as the interpretation and treatment decisions based on POCT results.

Standardization is an important aspect of successful POCT programs. Selecting one manufacturer for glucose meters and one kit for pregnancy testing simplifies quality management. With dozens of POCT sites and thousands of operators, increasing the number of different devices complicates the ability to manage the program and stresses available staff and resources. Validation, procedures, and training are required for each device for regulatory compliance. CAP also requires subscription to a proficiency-testing program. As the number of different devices multiplies, there are more procedures and training to maintain, and staff who rotate or test patients infrequently may not remember how to operate the test when required. For POCT, simplicity is fundamental. Fewer devices, limited testing sites, and reducing the number of operators simplify the management of testing and enhance compliance. With fewer devices and a smaller group of key POCT operators, it is easier to train staff and more economical to manage the program.

Method limitations

POCT methods are different from clinical laboratory methods and have their own biases and interferences. Clinical treatment protocols and pathways of care based on central laboratory methods may not necessarily be transferable to the same patient populations when using POCT results. As noted already, glucose meters may provide a faster glucose measurement, but the variability of the method may not allow the device to be used for diagnosis or screening of general patients, while central laboratory glucose results can be utilized for these situations. This is why clinical validation of a device on the intended population using routine staff operators at that location is so important before implementation of a test for patient care.

POCT devices are portable, so reagents and test strips may be exposed to varying environmental conditions compared with testing in a well-controlled laboratory. Visiting nurses travel to patient's homes, and tests can be exposed to cold in the winter and heat in the summer. Humidity can damage colorimetric test strips, like urine dipsticks. Altitude may also affect some tests that are sensitive to oxygen, like glucose and blood gas tests. Storage and environmental conditions are factors that need to be controlled to ensure quality test results. Tests and devices should be removed from vehicles when not in use to prevent freezing and overheating. Strip vials also need to be tightly recapped and tests stored in cool, dry places away from radiators and windows where the tests may be damaged from heat, light, and moisture.

Drugs, disease metabolites, and diet can affect POCT results in different ways compared with traditional laboratory methods. Some glucose strips are sensitive to salicylate, acetaminophen, and urea. Vitamin C in high doses can act as an electron carrier in POCT glucose reactions. Dietary constituents affect other POCT methods. Meat and peroxidase-containing vegetables, for instance, may give false-positive reactions in guaiac-based occult blood tests. Hydration status and drinking excess water can dilute urine, generating false-negative drug and pregnancy tests. Hemolysis of blood elevates potassium levels and may not be detected in a whole blood sample used for POCT compared with serum with laboratory analysis where hemolysis is evident as red color in the sample; so hemolysis can be overlooked with POCT, leading to falsely increased potassium results and the potential for misinterpretation and inappropriate clinical management.

Critical illness is another factor to consider with POCT. Patients presenting with extremes of physiology challenge the performance of POCT devices. Glucose meters originally developed for use in patient self-testing may perform differently when used in a hospital environment. Diabetic outpatients are ambulant and generally healthy compared with critical inpatients on oxygen therapy and various medications. Critically ill patients also have low hematocrits that can interfere with glucose methodologies. Manufacturers have been encouraged by the FDA to include a limitation in the meter package insert that "the performance of the test system has not been evaluated in the critically ill." While newer meters have received FDA approval for use in critically ill

populations, consumers are still warned of interferences with meter results when utilizing capillary fingerstick samples in patients with poor peripheral circulation. This limitation warns the consumer of the potential for interference in subpopulations of the critically ill that may show variations in results with certain sample types and conditions. Consumers should restrict use of glucose meters in order to adhere to package insert limitations, such as acceptable hematocrit and oxygen ranges, and prohibit use of capillary fingerstick samples from patients with poor peripheral circulation (i.e., diabetic ketoacidosis, nonketotic/hyperosmolar conditions, congestive heart failure stage IV, shock, and peripheral vascular disease) [15,16]. Blood can pool in the periphery and capillary samples in such patients may not reflect the physiology of central circulation. Currently, only one glucose meter has been FDA-cleared for use in critically ill patients using capillary fingerstick samples, but the analytical performance of this meter is significantly more variable than laboratory methods. Clinicians are cautioned to interpret POCT results with care in patients with rapidly changing physiologic conditions, particularly when analyzing capillary samples in critically ill patients despite FDA approval for use in these populations.

Use of venous or arterial samples may be a more acceptable alternative in patients with poor peripheral circulation, and capillary sampling is affected. Alternatively, hospitals may choose to use a different glucose methodology such as a blood gas analyzer or send a sample to the central laboratory [15]. However, delays in test turnaround may prohibit sending a sample to a central laboratory. Hospitals can also choose to use a glucose meter off-label by analyzing capillary samples despite the manufacturer warnings for critically ill patients. Caution should be taken, however, since off-label use shifts the CLIA-waived glucose meter to a high-complexity category requiring additional validation and documentation. CLIA high-complexity may further limit the staff who can perform testing, as hospitals frequently use patient care assistants, staff with only a high school education and on the job training, to perform glucose testing [15,17]. Under CLIA high complexity, staff with only a high school diploma would not be eligible to conduct glucose meter testing, and more advanced clinical staff, such as registered nurses or nurse practitioners, with additional laboratory training, experience, and licensure (in some states) may be required.

POCT employs unprocessed specimens. Whole blood is analyzed on POCT devices, whereas most laboratory methods require serum or plasma. Glucose, however, is unstable in whole blood. The specimen must be analyzed quickly after collection, or inhibitors of cellular glycolysis, like fluoride or oxalate, must be added to stabilize the glucose levels in the sample. Specimen additives, like fluoride and oxalate, can interfere with some POCT methods, so correlations of glucose meters against laboratory methods must quickly analyze a whole blood specimen on the glucose meter and either simultaneously collect a separate sample for laboratory analysis, or centrifuge and separate the remainder of the POCT specimen. Once separated from cells, glucose is stable in plasma/serum for several days refrigerated and longer if frozen. While use of whole blood for POCT and serum/plasma for laboratory analysis ensures appropriate specimen types for each methodology, the glucose concentration may not match between the two methods because of physiologic differences between whole blood and serum/plasma (Fig. 19.2).

Glucose distributes within the aqueous portion of blood. Since there is less water inside erythrocytes (73% water/27% lipids and protein) compared with the plasma portion of blood (93% water/7% lipids and protein), the glucose concentration will be lower inside red blood cells than plasma from the same patient specimen. The actual difference will depend on the hematocrit (concentration of erythrocytes in a sample). At a normal hematocrit of 45%, plasma is ~11% higher than whole blood. Patients with higher hematocrits will have greater difference between whole blood and plasma/serum glucose, while patients with lower hematocrits will have less difference. Method comparisons and validation of POCT devices must take into account these physiologic differences in specimen matrix when interpreting the agreement between different methods. To simplify interpretation, most

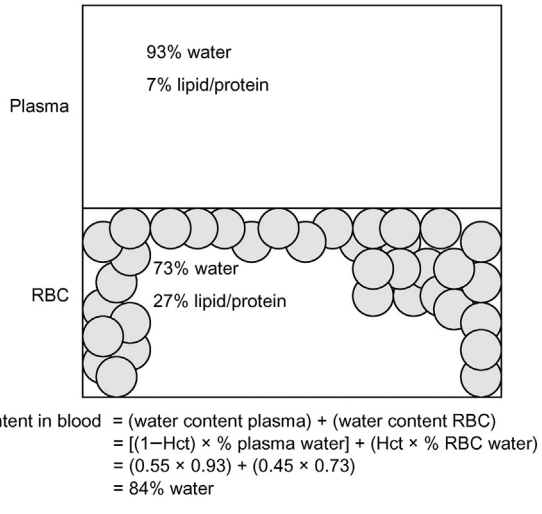


FIGURE 19.2 Whole blood to plasma differences. The physiologic water content in whole blood is a mixture of the plasma and erythrocyte red blood cell water and varies with the hematocrit or volume of red blood cells in blood. Analytes that follow water, like glucose, electrolytes, and some drugs, will show similar differences in concentration when measured in whole blood versus plasma.

manufacturers calibrate their glucose meters to read a plasma equivalent result despite accepting whole blood samples. Glucose meters should theoretically be comparable with central laboratory methods in patients with normal hematocrits. Differences in hematocrit are one reason why critically ill patients and neonatal populations with low and high hematocrits may show biases between the two methods, and why validation of POCT in the intended patient population is so important.

There are a variety of factors that can affect POCT differently from laboratory methods. These factors must be considered when implementing POCT or when interpreting the POCT literature. A device that works well on one medical unit and patient population may show different results in another patient population due to medication, diet, or operational differences between the two sites. Validation of POCT devices by staff operators on the intended patient populations will most likely uncover interferences and other potential problems before implementation and routine use of a device. When environmental effects or unexpected interferences are discovered, practices can be changed to control against test exposure or alternative devices can be selected, which do not experience the same interferences.

The potential for interference, operator, and environmental effects sets POCT apart from central laboratory methods. Results from POCT should be clearly separated from central laboratory test results in the patient's chart. As hospitals move toward electronic records, POCT result interfaces should distinguish the source of the test result for the ordering clinician. Clinical systems have the ability to overlay test results, so all glucose values or sodium values, for instance, can be viewed and trended together. From the clinician's perspective, seeing more information together on a computer screen saves time and prevents scrolling through result spreadsheets and multiple mouse clicks. However, differences in methodologies between POCT and central laboratory will cause results to jump around when plotted or trended together. This can lead to clinical confusion and chasing results where treatment is based on one result, but the next result comes back too low. Based on this, therapy is readjusted; then the next result is too high, so therapy is adjusted again. Combining results from multiple methods can lead to such clinically confusing trends, overtreatment, and frequent changes in treatment that would not occur if the individual methods were separated in the patient's medical record.

Other distinctions that should be identified in the medical record are calculated and measured results. Some blood gas analyzers measure hematocrit by conductivity and calculate hemoglobin by a common conversion (hemoglobin = hematocrit/3), while other analyzers measure hemoglobin spectrophotometrically and calculate hematocrit. Conductivity hematocrit is sensitive to protein concentration and can be biased in surgical and trauma patients who receive transfusions and volume replacement. So POCT results for hemoglobin and hematocrit should specify not only that result was performed point-of-care, but whether the result was calculated or measured. These are only a few of the interface challenges to consider because of the unique method differences and biases that can occur with POCT in some patient populations.

Risk management

POCT systems are not perfect. All devices have limitations. When used outside manufacturer recommendations or used improperly, incorrect results can be generated. Errors can be more likely with POCT because of the number of operators involved, the volume of tests being conducted, and the various locations where testing is performed. POCT is often delegated to the lowest level of staff who have minimal training and experience. Staff shortcuts, differences in operator technique, and failure to appreciate patient and preanalytic factors can all compromise test quality. POCT is chosen over laboratory testing, because it is rapid and convenient. POCT can be performed at a variety of locations with the opportunity for the device and the reagents to be exposed to heat, cold, and humidity. The number of operators, tests, and locations multiplies the complexity of managing the testing process in a quality manner. Environmental exposure, sample, and operator errors can all lead to incorrect results.

Risk management can reduce errors. Manufacturers employ risk management when developing new methods in order to document that common hazards or the potential for errors have been addressed by the device engineering and that the limitations have been outlined in the package insert and instructions for use. The FDA considers the manufacturer risk management plan when reviewing the device for market approval. This plan outlines how the device is intended to be used (use-case scenario) and how the manufacturer addresses or detects potential misuse. With POCT, a device may be brought to the patient's bedside for testing, or alternatively a sample may be collected and brought to a device in a spare utility room on the nursing unit for testing. Moving a sample away from the patient presents the potential to mix-up samples with other patients tested in that utility room, unless the sample is labeled appropriately. These two use-case scenarios present different opportunities for particular errors, such as specimen mix-up. Risk management examines each step in the testing process for weaknesses where errors may occur and defines actions that may be required to reduce, prevent, or detect those errors before a result is released and action is taken by clinicians. So

laboratories can learn to reduce their chance for test errors by adopting industrial risk management principles.

Testing liquid QC specimens in a manner like patient specimens have historically been utilized to prove the ability of a test system to achieve a test result with the expected level of quality; however, liquid controls have limitations. With POCT, analyzing a QC sample consumes a single-use test kit, and there is no guarantee that the next kit will perform comparably. Based on this, manufacturers have developed a number of control processes engineered into the test or actions required to be performed by the laboratory to achieve quality results. Newer devices have clot and bubble detection, volume detection to ensure adequate sample was applied, and internal controls to prove the viability of reagents such as internal controls on pregnancy, HIV, and drug tests. Reagents are being shipped with a temperature monitor to ensure the package has not been exposed to heat or cold temperatures outside of manufacturer specifications during shipping. Staff are expected to check the temperature monitor on receipt of each shipment and verify acceptability before attempting to place the kits in use for patient testing. Each POCT device and test kit is unique and may present different opportunities for error depending on how the device is used by the consumer. Applying risk management to trace the sample through the testing process can reveal hazards or potential sources for error and the actions that are required to prevent and detect those errors. CLSI has published a guideline for assessing the risk of errors from implementing unit use testing devices, titled "Quality Management for Unit-Use Testing" [18]. This document advocates the use of an error matrix to list and prioritize the frequency and severity of potential errors associated with testing devices. This guideline also emphasizes a shared responsibility between the manufacturer and the consumer of the testing device in developing appropriate control strategies to manage errors. Another CLSI document, EP 23-A, titled "Laboratory Quality Control Based on Risk Management" [19], outlines a strategy to develop an IQCP for the laboratory test based on risk management. An IQCP basically summarizes many of the quality activities a laboratory is already doing in a more formal manner. The IQCP is discussed in further detail in Chapter 5, Quality control. Manufacturers have been conducting risk assessments for years and must submit their data for FDA review when seeking market approval of new tests. CLSI EP23 and the IQCP introduce these industrial risk management principles into the clinical laboratory.

Summary

POCT is an increasingly popular means of delivering faster test results closer to the patient. As the number of devices and menu of tests expands, the POCT market can be expected to grow. The major concern with POCT is providing quality test results with multiple devices, at dozens of sites by thousands of operators. Management of POCT, therefore, requires organization. POCT management can best be organized by forming a POCT committee to supervise new test requests and set policies, standardizing and minimizing the number of devices, planning implementation through validation of devices and training of operators, and providing consistent laboratory oversight to comply with regulatory requirements. POCT is a different method and may not be freely interchangeable with laboratory results in clinical protocols. Each method has unique interferences, biases, and imprecision that can lead to disparate results in certain patient populations. The potential for method differences stresses the need to validate devices at the intended clinical sites by using clinical staff operators to test typical patients before the decision to implement a device for patient care. Newer devices are computerized and can automatically collect and transfer the test results and QC data necessary to document regulatory compliance. Interfaces of POCT results should distinguish POCT results from laboratory results in order to prevent clinical confusion that can occur from overlaying results with unique biases in certain patient populations and clinical situations. Many newer POCT devices also offer improvements in detecting common errors through internal control processes. Risk management offers a means for laboratory directors to determine the required actions and optimal control strategies to minimize risk of errors for any specific device in their healthcare setting and use-case scenarios. POCT has the potential to provide faster results and improve patient outcomes through quicker therapeutic intervention provided that the clinician understands the method limitations and when to use the devices in patient care appropriately. POCT is thus a complex system with many factors to consider and manage to obtain optimal benefits.

References

[1] J. Greyson, Quality control in patient self-monitoring of blood glucose, Diabet. Care 16 (1993) 1306–1308.

[2] W.A. Rutala, V.A. Kennedy, H.B. Loflin, F.A. Sarrubi Jr., Serratia marcescens nosocomial infection of the urinary tract associated with urine measuring containers and urinometers, Am. J. Med. 70 (1981) 659–663.

[3] R. Webb, M. Currier, J. Weir, K.M. McNeill, E. Bancroft, D. Dassey, et al., Transmission of hepatitis B virus among persons undergoing blood glucose monitoring in long-term care facilities—Mississippi, North Carolina and Los Angeles County, California, 2003–2004, MMWR Morb. Mortal. Wkly. Rep. 54 (2005) 220–223.

[4] Department of Health and Human Services, Office of Inspector General. Enrollment and certification processes in the Clinical Laboratory Improvement Amendments Program. OEI-05−0000251. Chicago, IL: Office of Evaluation and Inspections, 2001, p. 18. <http://oig.hhs.gov/oei/reports/oei-05-00-00251.pdf> (accessed 05.03.2020).

[5] CMS. CMS certificate of waiver and provider performed microscopy procedures pilot project. Baltimore, MD: CMS, 2003, p. 30. <http://www.cms.hhs.gov/CLIA/downloads/ppmpfr2001.pdf> (accessed 05.03.2020).

[6] Clinical Laboratory Standards Institute, Point-of-Care Connectivity: Approved Standard POCT1-A2, CLSI, Wayne, PA, 2006. 306pp.

[7] G.J. Kost (Ed.), Principles & Practice of Point of Care Testing, Lippincott, Williams & Wilkins, Philadelphia, PA, 2002. 654pp.

[8] J.H. Nichols (Ed.), Point-of-Care Testing: Performance Improvement and Evidence-Based Outcomes, Marcel-Dekker, New York, NY, 2003. 500pp.

[9] E.T. Chen, J.H. Nichols, S.H. Duh, G. Hortin, Performance evaluation of blood glucose monitoring devices, Diabetes Technol. Ther. 5 (2003) 749−768.

[10] D.B. Sacks, D.E. Bruns, D.E. Goldstein, N.K. Maclaren, J.M. McDonald, M. Parrott, Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus, Clin. Chem. 48 (2002) 436−472.

[11] D.B. Sacks, M. Arnold, G.L. Bakris, D.E. Bruns, A.R. Horvath, M. S. Kirkman, et al., Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus: Updated draft guidelines, National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines, 2011. 120pp. <http://www.aacc.org/science-and-practice/practice-guidelines/diabetes-mellitus> (accessed 05.03.2020).

[12] Clinical Laboratory Standards Institute, Point-of-Care Blood Glucose Testing In Acute and Chronic Care Facilities; Approved Guideline − Third Edition POCT12-A3, CLSI, Wayne, PA, 2013. 49pp.

[13] Centers for Disease Control, Laboratory Practice Assessment Br, Division of Laboratory Systems, Public Health Practice Program Office. Clinical laboratory performance on proficiency testing samples − United States 1994. MMWR Morb Mortal Wkly Rep 1996;45:193−6.

[14] R. Belsey, J.I. Morrison, K.J. Whitlow, D.M. Baer, S. Nelson, D. F. Hardwick, Managing bedside glucose testing in the hospital, JAMA 12 (1987) 1634−1638.

[15] Clinical Laboratory Standards Institute, Use of glucose meters for critically ill patients. POCT17, CLSI, Wayne, PA, 2016. 15 pp.

[16] IFCC Working Group GMECC. How should glucose meters be evaluated for critical care? Milan, Italy: International Federation of Clinical Chemistry and Laboratory Medicine, 2018, p. 48. <http://www.ifcc.org/ifcc-education-division/emd-committees/c-poct/wg-gmecc/> (accessed 05.03.2020)

[17] US Department of Health and Human Services, Medicare, Medicaid, and CLIA programs; laboratory requirements relating to quality systems and certain personnel qualifications; final rule. 42CFR Part 493, Fed. Regist. 68 (2003) 3639−3714.

[18] Clinical Laboratory Standards Institute, Quality Management for Unit-Use Testing: Approved Guideline. EP18-A, CLSI, Wayne, PA, 2002. 38pp.

[19] Clinical Laboratory Standards Institute, Laboratory Quality Control Based On Risk Management: Approved Guideline EP23-A, CLSI, Wayne PA, 2011. 73pp.

Self-assessment questions

1. What is the definition of point-of-care testing?

a. Testing conducted close to the site of patient care
b. Bedside testing
c. Portable testing in an ambulance or helicopter
d. All of the above

2. Which of the following tests are available from POCT devices?

a. Growth hormone
b. Dihydrotestosterone
c. HIV antibody
d. Renin

3. What are the advantages of POCT?

a. Larger sample
b. Faster turnaround time for test results
c. Less expensive than core lab testing
d. More precise than core lab instrumentation

4. Which regulations apply to POCT devices?

a. Federal CLIA regulations
b. State law
c. Private accreditation agency standards like The Joint Commission
d. All of the above

5. Which of the following are true?

a. POCT results can lead to wrong treatment, and care should be taken to guarantee quality.
b. POCT devices are so simple that they pose no risk to the patient if incorrectly performed or a wrong answer is generated.
c. POCT results are used only for screening patients, and results are always confirmed by a lab method.
d. Clinicians can purchase POCT devices from a local pharmacy without having to comply with laboratory regulations if the tests are used only on their own patients in an office practice.

6. The DHHS enforces CLIA standards in which way(s)?

a. Limiting the physician's medical license to practice
b. Stopping reimbursement from Medicare and Medicaid
c. Closing the laboratory or physician's office performing testing.
d. All of the above

7. POCT device connectivity and the POCT1 standard are important for what reason(s)?

a. Allows multiple devices to connect to a single data manager and share an interface with a laboratory or HIS
b. Allows collection and review of data from multiple sites
c. Facilitates implementation of new POCT devices through plug-and-play connections
d. All of the above

8. Limitations of POCT devices include which of the following?

a. Ability of physicians to rapidly treat patients
b. Environmental and drug interferences with rapid methods
c. Ease of use and training of the operator
d. Small sample volumes

9. Which of the following is a QC process?

a. A whole blood control above the medical decision limit for a test
b. A colored disk to check the optics of a spectrophotometric device
c. A separate control line that develops with each test on a pregnancy strip test
d. All of the above

10. What is risk management?

a. A requirement for FDA approval of new devices
b. An industrial process for defining potential errors in the use of a device
c. A means for clinical laboratories to prevent and detect errors
d. All of the above

Answers

1. d
2. c
3. b
4. d
5. a
6. b
7. d
8. b
9. d
10. d

Chapter 20

Applications of molecular techniques in the clinical laboratory

Allison B. Chambliss[1] and Mark A. Marzinke[2]

[1]*Department of Pathology, Keck School of Medicine of USC, Los Angeles, CA, United States,* [2]*Departments of Pathology and Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, United States*

Learning objectives

After reviewing this chapter, the reader will be able to:

- Describe molecular techniques used in the evaluation of cancers.
- Explain the role of nucleic acid testing in the identification of common infectious pathogens.
- Identify the influence of genetics in the coagulation cascade.
- Understand regulatory considerations associated with molecular testing.

Advances in genomic technologies have exponentially increased knowledge of human genetics and have led to the identification of numerous inherited and acquired genetic variants. Characterization of such genetic variants is critical to the differential diagnosis of numerous disease states, the selection of treatment regimens, and the monitoring of treatment responses. Traditionally, variant detection has been performed in a targeted manner, thus limiting analysis to a predefined selection of sequence changes associated with a clinical phenotype. However, more recently, multiplexed and next-generation sequencing (NGS) technologies have allowed for broader exome- or genome-wide approaches to facilitate inherited or acquired variant identification.

Molecular methodologies have been utilized in the characterization of both native and foreign genomic sequences. For instance, molecular platforms are commonly employed in constitutional genomic testing, which can be used to characterize congenital or developmental disorders. Molecular applications to the diagnosis of inborn errors of metabolism are discussed in more detail in Chapter 48, Inborn errors of metabolism and newborn screening, of this book. In addition, molecular testing is widely used in the identification of infectious pathogens and in the characterization of viral drug resistance. The central backbone of molecular platforms is the polymerase chain reaction (PCR)-mediated amplification of native, foreign, or acquired DNA regions. PCR amplification may occur in a targeted, untargeted, or multiplexed manner. Amplicon quantification or detection may be achieved using a variety of oligonucleotide or sequencing-based approaches; a review of nucleic acid probes and PCR technology is found in Chapter 13, Nucleic acid analysis in the clinical laboratory. This chapter will discuss the current applications of molecular analysis as well as the clinical utilities, limitations, and regulatory considerations associated with these analytical approaches.

Molecular diagnostics

Somatic gene variants

Cancer progression may be associated with the acquisition of genetic variants, which may result in the alteration of DNA replication, cellular proliferation, or differentiation. Such somatic variants are typically identified in the laboratory via analysis of DNA extracted from formalin-fixed, paraffin-embedded tissue. As tissue biopsies are heterogeneous, cancerous regions often contain populations of various distinct clonal populations of tumor cells as well as "contamination" by normal cells. Thus pathological findings and postreview processing of tissue samples are critical to the identification of acquired genetic variants. Further, clinical assays must be sensitive enough to distinguish the targeted variant among a heavy background of nonvariant DNA.

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00020-X

There are a number of transcription factors and signaling pathways that control molecular processes within cells. For example, the MAP kinase/ERK pathway, also known as the Ras−Raf−MEK−ERK pathway, is a signaling cascade that modulates protein translation and cellular division. Activation of the Ras−Raf−MEK−ERK pathway is controlled via ligand binding to the signaling epidermal growth factor receptor (EGFR) molecule. Ligand binding results in intracellular EGFR phosphorylation and downstream activation of adapter molecules; it is through this process that the Ras protein can bind the activator guanosine triphosphate (GTP) and initiate the aforementioned signaling cascade. Genetic variations can lead to constitutive pathway activation, thus resulting in uncontrolled cellular proliferation; dysregulation of the cascade is a key factor in cancer progression. Therapies directed toward the EGFR molecule, such as tyrosine kinase inhibitors and monoclonal antibodies, have demonstrated promising activity in the treatment of several cancers. For instance, EGFR-targeted therapies are widely employed in the treatment of colorectal cancers, nonsmall-cell lung cancers, and squamous cell carcinomas of the head and neck.

Specific variations within the genes located downstream of EGFR can render the pathway constitutively active even if EGFR itself is inhibited. Consequently, tumors with these variants are unresponsive to EGFR-targeted therapies. For instance, the acquired V600E mutation of the *BRAF* proto-oncogene results in the substitution of glutamic acid for valine at position 600 within the protein chain, resulting in a constitutively active BRAF signaling molecule. There is a prevalence of approximately 10% for the *BRAF* V600E mutation in colorectal cancers, and the somatic variant predicts resistance to anti-EGFR monoclonal antibody therapies such as cetuximab and panitumamab. Further, there is an estimated frequency of 45% for this variant in papillary thyroid cancers (PTCs). PTCs carrying the *BRAF* V600E variant are associated with a more aggressive phenotype that may include extrathyroidal invasion, lymph node metastases, and tumor recurrence [1]. Similarly, variants in the *KRAS* oncogene downstream of EGFR are implicated in carcinomas of the lung and colon. As many as 40% of colorectal cancers involve mutations in *KRAS*, with approximately 90% of variants found in codons 12 and 13 [2]. Thus guidelines from the National Comprehensive Cancer Network include tissue genotyping for *KRAS* variants in all patients with metastatic colorectal cancer. Table 20.1 lists other representative oncogenic variants commonly assessed in the monitoring and treatment of various cancers.

Variants within the EGFR pathway were previously detected by direct sequencing methods such as Sanger sequencing or pyrosequencing for single-gene analysis; however, detection of these variants via NGS platforms is becoming more common [3]. As variants must be distinguishable from the background of normal genetic sequences, specific considerations should be taken into account with regard to the analytical limits of detection (LODs) of the methodological technologies used. The analytical LODs vary between the methods and the laboratories but typically fall within 5%−30%, which refers to the percentage of DNA from the extracted material that carries the genetic aberration. Thus the variant of interest

TABLE 20.1 Representative oncogenic mutations commonly assessed in cancer.

Oncogene	Chromosome location	Associated protein function	Type of mutation	Representative associated cancers
HER2/*neu* (*ERBB2*)	17q12	Receptor tyrosine kinase	Gene amplification	Breast, ovarian, gastric, colon
Ras family: *HRAS* *KRAS* *NRAS*	11p15.5 12p12.1 1p13.2	Signal transduction (GTP/GDP binding)	Point mutations	Pancreatic, colorectal, lung, melanoma, bladder
BRAF	7q34	Signal transduction	Point mutations	Colorectal, papillary thyroid, melanoma
c-myc	8q24.21	Transcription factor	Translocation	Lymphoma, leukemia
N-myc	2p24.3	Transcription factor	Gene amplification	Neuroblastoma
PIK3CA	3q26.3	Phosphorylation	Point mutations	Colorectal, ovarian, breast

will only be detected if its frequency among nonvariant tissue DNA meets or exceeds the assay LOD. For example, general LODs have been estimated as 15%−20% for Sanger sequencing, 5% for pyrosequencing, and 10% for melting curve analysis [4]. In addition to testing of the tumor tissue itself, genetic variants in genes such as *KRAS* may be detected in cell-free DNA (cfDNA) obtained from plasma samples. Analysis of cfDNA is discussed later in this chapter.

Loss of heterozygosity

Loss of heterozygosity (LOH) occurs when a somatic cell contains only one of the two traditional alleles at a genetic locus. LOH may be the result of the failure of homologous chromosomes to separate properly during mitosis (nondisjunction), recombination segregation of homologous chromosomes before separation, or the deletion of a chromosomal segment. Cancer may develop if the missing chromosomal region includes the genomic information encoding a tumor suppressor gene. Further, LOH may also present as copy neutral by uniparental disomy, which is the consequence of maternal or paternal chromosomal duplication with the concurrent loss of the other allele (the Knudson two-hit hypothesis) [5]. In these cases, cancer may arise if the nonfunctional allele of a tumor suppressor gene is inherited. LOH has been reported in numerous cancer types, such as melanoma and cancers of the breast, lung, prostate, and bladder. LOH is considered the most frequent molecular genetic alteration observed in human cancers [6]. Further, when present in the germline, uniparental disomy may be associated with such congenital diseases as Prader−Willi or Angelman syndromes.

Analysis of microsatellite targets is the primary methodological approach employed for LOH detection. Microsatellites are tandem repeats of one to six nucleotides scattered at a frequency of approximately 3% across the genome [7]. These repeat elements are unstable and are prone to mispairing during DNA replication. Wild-type mismatch repair (MMR) genes typically correct these replication errors, resulting in microsatellite stabilization in the genome of healthy patients. However, variants in MMR genes have been implicated in the pathogenesis of several cancers and result in variability with respect to microsatellite length, causing microsatellite instability (MSI). Germline alterations in MMR genes are the cause of hereditary nonpolyposis colon cancer (Lynch syndrome), which account for approximately 5% of all colon cancer cases. Thus the detection of MSI may reflect cancer-causing MMR variants and can, therefore, be used to identify the presence of LOH or allelic imbalances.

A targeted microsatellite can be analyzed via PCR amplification using primers flanking the targeted tandem-repeat region. Products are subsequently separated by size via gel or capillary electrophoresis. The absolute size of each microsatellite may be determined using an internal size standard, which employs an additional dye to provide accurate allelic size determination and eliminates the possibility of run-to-run electrophoretic variability. MSI is defined as a change of allelic length due to the insertion or deletion of repeating units. Thus by comparing nontumorigenic control DNA (i.e., from lymphocytes) to DNA isolated from solid tumor tissue from the same patient, MSI may be detected by shifts in allelic size from the normal control (Fig. 20.1). MSI may also be detected in cfDNA from genetic testing of blood. Importantly, microsatellite typing can detect copy-number neutral LOH, which is not possible with traditional karyotyping techniques or fluorescence in situ hybridization (FISH). The latter is a common technique used to visualize specific DNA sequences on chromosomes. MSI is associated with endometrial and bladder cancers, among others. Additional

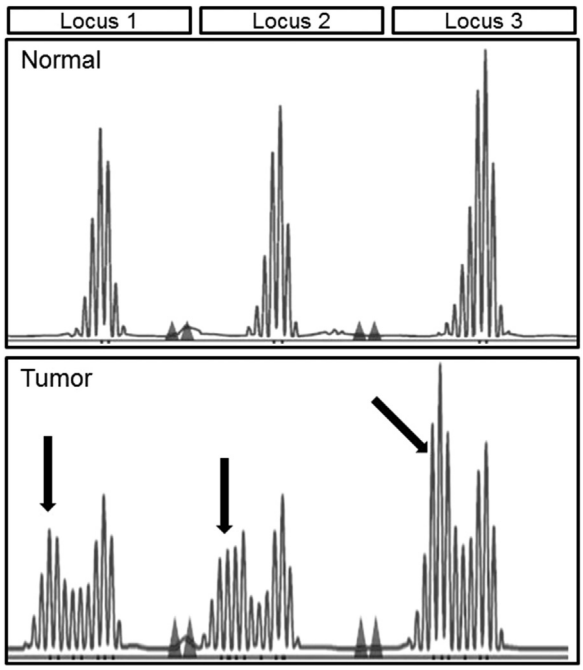


FIGURE 20.1 Microsatellite instability. Microsatellite instability, associated with many cancers, occurs when mismatch repair genes become mutated and cannot repair DNA mispairing errors. Resultant affected sequences from tumor tissue can be identified after targeted polymerase chain reaction amplification by a change in allele base pair length when compared with a nontumorous DNA control. This example for three targeted genome loci plots a frequency histogram of measured polymerase chain reaction product sizes measured by capillary electrophoresis (relative frequency on *y*-axis, allele size on *x*-axis). The tumor tissue DNA (*bottom panels*) shows additional product sizes, indicated by the black arrows, which are not present in the normal control DNA (*top panels*), indicating microsatellite instability. Data were collected via a microsatellite instability multiplex kit (Promega) and analyzed with GeneMapper software (Applied Biosystems/Thermo Fisher Scientific).

applications of microsatellite profiling outside of cancer include determination of parental origin of new mutations as well as engraftment testing by comparison of pre- and posttransplant samples of donor and recipient DNA.

Single nucleotide polymorphism (SNP) arrays are another common tool to assess LOH; these arrays are used for the genome-wide detection of copy-neutral LOH. This type of analysis involves assessment of a particular genotype (represented generally as homozygous AA, heterozygous AB, or homozygous BB) via single-base extension using labeled oligonucleotide probes at hundreds of thousands of known SNP locations simultaneously throughout the genome. SNP arrays may provide greater insight into the characterization of leukemias and other hematologic malignancies with characteristic chromosomal deletions. SNP arrays are also a common approach utilized for the assessment of LOH on chromosome arms 1p and 19q; the presence of LOH at these locations is diagnostic of oligodendroglioma in patients presenting with brain tumors. Similarly, LOH at chromosomes 1p and 16q is associated with a poor prognosis in patients with Wilms tumor, a renal malignancy observed in children [8]. Furthermore, upon detection of ultrasound abnormalities during fetal development, whole-genome SNP arrays provide a more rapid and high-resolution analysis of unbalanced germline chromosomal abnormalities in the fetus than traditional karyotyping.

Circulating tumor cells and cell-free DNA

Tumor cells enter the circulatory system when they break from primary tumors of epithelial origin. In cancer, circulating tumor cells (CTCs) may exist at a frequency of one per 5−10 million blood cells and may be collected and enriched for quantification. The Menarini Silicon Biosystems CellSearch platform, the only FDA-cleared test for the assessment of CTC at the time of this publication, is approved for use in patients with breast, prostate, and colorectal cancers. The assay uses a combination of immunomagnetic cell labeling and automated digital microscopy to negatively select for and quantify isolated tumor cells in 7.5 mL blood [9]. While tissue biopsies are risky and invasive, genetic testing from blood requires only a blood draw and is thus an attractive alternative for characterizing or monitoring the aforementioned cancers. Increased CTC counts are correlated with poorer prognosis and shorter survival rates in many cancers [10].

Tumor-derived extracellular DNA provides another target for the detection of characteristic genetic or epigenetic mutations. All patients, healthy and diseased, possess some portion of extracellular or cfDNA that is released upon apoptosis and circulates unbound from leukocytes or erythrocytes. The presence of such cfDNA was first described in 1948 by Mandel and Metais [11]. However, it was not until the 1990s that their discoveries were further extrapolated with the use of PCR to demonstrate that oncogenic variants could be detected in the plasma of cancer patients [12,13]. Presently, detection of variants in both proto-oncogenes and tumor suppressor genes, as well as LOH analysis (i.e., microsatellites, genetic aberrations), can be performed using plasma. Because plasma contains cfDNA both of cancerous and noncancerous tissue origin, appropriately specific and sensitive methods are required to identify malignant genetic material. Initial methods implemented for cfDNA analysis included allele-specific PCR amplification with direct sequencing; however, such methods often lacked sensitivity to detect the rare amounts of cfDNA in the plasma. More recently, single-molecule PCR, also known as digital PCR (dPCR), has been utilized due to its ability to detect a variant present in a very small fraction of the whole specimen. The principles behind dPCR are discussed later in this chapter.

In addition to cancer applications (i.e., monitoring of tumor burden, minimal residual disease, and molecular resistance), cfDNA has also revolutionized the field of fetal medicine. Inspired by the discovery of tumor-associated DNA in plasma, Lo and colleagues set out to determine whether, in the same manner, fetal DNA could be found in the plasma of the mother. In 1997 they discovered that a sequence from the Y chromosome was detectable in the plasma of pregnant women carrying male fetuses [14]. The fraction of fetal DNA in maternal plasma may approach 10% in the first and second trimesters and 20% in the third trimester [15]. Analysis of fetal DNA by liquid biopsy overcomes the risks and invasiveness of traditional sampling methods such as amniocentesis and chorionic villus sampling (CVS). The first clinical application of DNA-based noninvasive prenatal diagnosis came with the demonstration that fetal rhesus D blood group status could be identified in maternal plasma via real-time PCR (qPCR) [16]. Other applications include the detection of paternally inherited variants associated with thalassemias and cystic fibrosis. dPCR techniques can distinguish low abundance fetal genomic sequence from the primarily maternal genomic background. However, the detection of fetal chromosomal aneuploidies, such as trisomy 21, has proven more challenging, because it requires specific differentiation between the fetal and maternal chromosomes. One approach to circumvent this issue involves the assessment of an epigenetic allelic ratio via the analysis of differential DNA methylation between the maternal and fetal chromosomes. Currently, this approach is considered a laboratory-developed test.

Massively parallel sequencing (MPS) has made it possible to count millions to billions of DNA molecules in order to detect accurately the presence of a trisomic fetus.

Assays for the detection of trisomy 21, 13, and 18 by noninvasive methods are now commercially available. Importantly, the American College of Obstetricians and Gynecologists currently recommends that cfDNA testing be considered only for screening purposes, and all abnormal results should be confirmed by diagnostic testing via amniocentesis or CVS [17]. While cfDNA testing eliminates the invasiveness and risks of traditional prenatal screening, there also exist ethical concerns, as the ease of testing may encourage more those mothers with "increased risk" to undergo testing.

Genetic variants in hemostasis

The use of molecular diagnostic techniques in hematology laboratories has increased with both the improvement of analytical technologies as well as the identification of genetic factors affecting coagulation. The coagulation system functions to control bleeding through the generation of blood clots. The process is complex and dynamic and involves the activation, adhesion, and aggregation of platelets to generate ultimately cross-linked fibrin. Regulators of the involved pathways include several coagulation factors, which are generally serine protease (PR) enzymes or transglutaminases. Molecular assays serve to assess function along many various steps of the coagulation cascade.

Prothrombin polymorphisms

Prothrombin, also known as factor II, is the vitamin-K-dependent precursor to thrombin (factor IIa) at the culmination of the coagulation cascade. A well-characterized variant exists within the prothrombin gene on chromosome 11p11-q12 that is defined by a guanine-to-adenine (G > A) substitution at nucleotide 20210 in the 3′ untranslated region. The primary consequence of this genetic aberration is increased prothrombin accumulation with a resultant two- to threefold increased risk of venous thromboembolism. The prevalence of this variant is $<1\%-6\%$, and the frequency is dependent on ethnic background [18]. Of note, plasma prothrombin levels alone are insufficient to identify carriers of the defect due to the large inter- and intraindividual variabilities across populations. Thus genetic analysis is necessary.

PCR-based assays for prothrombin variant evaluation are widely performed using DNA obtained from whole blood; combined restriction digestion and agarose gel electrophoresis may be used for aberration detection. Techniques that are more modern include beacon-based or fluorescent probes using sequence-specific hybridization as well as high-resolution melting curve analysis. With regard to the melting curve analysis technique, PCR amplification precedes a gradual temperature increase, which results in the dissociation of the amplified product at a specific melting temperature dependent on the base pair sequence of the amplicon. The melting temperature is defined as the point at which the fluorescent dye (which is intercalated into double-stranded DNA) signal decreases following product dissociation. The G > A variant decreases the amplicon melting temperature from a wild-type control and readily allows for variant identification in the sample. Of note, other less-prevalent polymorphisms of the prothrombin gene have been described, including C20209T and A19911G. These variants are not as well characterized as G20210A. However, both polymorphisms are also thought to increase plasma prothrombin levels and heighten risk of venous thromboembolism.

Factor V Leiden polymorphism

Upstream of factor II in the coagulation cascade, activated protein C (APC) functions to degrade factors V and VIII, thus acting as a natural anticoagulant to balance thrombin production. A genetic polymorphism causing an arginine-to-glutamine substitution at nucleotide 1691 within factor V gene, which is located on chromosome 1q23, leads to a defect in the APC cleavage site. This amino acid substitution causes decreased factor V susceptibility to inactivation by APC and is associated with resistance to APC activity. The abnormal factor V protein is termed factor V Leiden (FVL) and is the most common cause of inherited thrombophilia in Caucasians, with as much as 5% of the population possessing heterozygosity for this polymorphism. Further, prevalence may approach 15% in some European and Middle Eastern countries, while the variant is typically not observed in African and Asian populations. The FVL point mutation may be detected using similar analytical approaches as the aforementioned prothrombin G20210A variant. Commercial assays often utilize PCR followed by target-specific hybridization with reporter fluorophores, such as a TaqMan probe-based approach; this methodology is described in more detail in Chapter 13, Nucleic acid analysis in the clinical laboratory. Based on measured fluorescent signal, allelic discrimination data analysis approaches determine whether the assayed sequence is wild type (or reference), heterozygous, or homozygous for the targeted variant.

Methods are also available to detect both the G20210A prothrombin variant and the FVL variant in the same reaction. This is achieved by multiplexed PCR of both affected genetic regions followed by similar target-specific hybridization methods. It is estimated that approximately 1 in 1000 Americans are doubly heterozygous for both FVL and G20210A, significantly increasing the risk of venous thromboembolism 20-fold over the general population [19]. For both FVL and

G20210A, screening in asymptomatic patients is typically limited to those with a strong family history of venous thromboembolism.

Nonmolecular approaches to identify the abnormal FVL factor are functional in methodological design. For example, APC resistance assays consist of diluting patient plasma into factor V-deficient plasma for the evaluation of activated partial thromboplastin time. Functional APC resistance assays may be used as a screen for FVL; however, positive results should be confirmed by genetic testing to confirm the diagnosis and determine hetero- or homozygosity in order to stratify thrombotic risk [20].

Methylenetetrahydrofolate reductase polymorphism

The 5,10-methylenetetrahydrofolate reductase (MTHFR) enzyme is responsible for the conversion of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate in the metabolism of folate. A mutation in the gene encoding for MTHFR represents the most common form of genetic hyperhomocysteinemia. This *MTHFR* polymorphism is characterized by a cytosine-to-thymine substitution at nucleotide 677, leading to an amino acid replacement of valine for alanine at codon 222. The mutation variably reduces MTHFR enzyme activity by rendering the enzyme more thermolabile. Variants may occur with frequencies as high as 40% in the general population [21]. Affected individuals may present with increased risk of venous thromboembolism and cardiovascular disease. Targeted genotyping assays for *MTHFR* may aid in the differential diagnosis of megaloblastic anemia, especially in the background of folate deficiency and elevated plasma homocysteine concentrations. Testing may also be indicated for individuals on medications known to affect folate metabolism (i.e., methotrexate). Mutational status is often measured using molecular probe-based approaches or melting curve analysis.

Nucleic acid testing in pathogen detection and monitoring

Nucleic acid testing has transformed the management of pathogen detection and monitoring in infectious diseases. In contrast to traditional culture methods, rapid molecular methods facilitate expedited result reporting and provision of clinical decision support. Further, in some cases, quantification of pathogen nucleic acids provides a means of monitoring infection progression in addition to monitoring response to therapy. While infectious diseases commonly assessed in the clinical laboratory are described in more detail in Chapter 54, Infectious diseases, of this book, this section will focus on nucleic acid testing for pathogens for which molecular technologies play a large role.

Hepatitis B virus

Hepatitis B virus (HBV) is a circular, partially double-stranded DNA virus from the hepadna family. HBV is the most common cause of acute hepatitis globally and is the most common form of chronic viral infection. The HBV surface antigen and other surface protein forms contain various molecular subdeterminants, which distinguish 10 genotypes (A-J) and >30 subtypes of the virus. In the United States, the A, B, and C genotypes are most prevalent. The complicated structure of HBV includes several enzymatically active sites that make viral reproduction error-prone, leading to a high frequency of viral mutations and yielding challenges in vaccination, diagnosis, and treatment. For example, a relatively common G > A point mutation at nucleotide 1896 inserts a stop codon in the place of tryptophan, thereby preventing the translation of a functional hepatitis B e antigen (HBeAg), rendering affected individuals' infections undetectable when using HBeAg immunoassay testing. This variant may be present in over 70% of patients in the United States with HBV type D [22]. Similarly, variants in the "a" determinant region of the HBV surface antigen (HBsAg), which is involved in the binding of antibodies against the antigen, may cause patients to test falsely negative when using the HBsAg immunoassay, which is the most widely used marker to assess an active HBV infection. Quantification of HBV DNA is critical to the diagnosis and management of infection, especially when serology testing is inconclusive.

HBV DNA assays are available for monitoring treatment responses and viral genotyping. For DNA analysis, most assays use qPCR techniques with LODs as low as 10 IU/mL. International units are used rather than the traditionally reported copies/mL due to the establishment of international reference material from the World Health Organization [23]. However, due to various methodologies used for HBV DNA quantification, conversion between IU/mL and copies/mL tends to differ amongst assays. Roughly, 1 IU/mL equates to approximately 5 copies/mL, though PCR-based quantification assays have demonstrated higher conversion factors than branched DNA (bDNA) assays. bDNA assays consist of a series of hybridization steps that result in a branched "sandwich" complex of probes and target sequence to amplify signal [24]. Accordingly, interpretation of viral load is somewhat controversial, and patients should ideally be monitored with the same assay. Some guidelines suggest a cutoff value of 20,000 IU/mL to diagnose the virus. However, a lower threshold is necessary for HBeAg negative individuals, such as those with the G1896A variant, who naturally have lower viral loads despite an aggressive disease course when chronically affected [25]. Thus serial monitoring of DNA levels is

considered more valuable than a cutoff threshold when determining treatment regimens. HBV genotyping is not widely performed, as further studies regarding the role of genotyping in the guidance of clinical decisions are necessary before it can be recommended for routine practice.

Hepatitis C virus

Hepatitis C virus (HCV) is a single-stranded RNA flavivirus that is transmitted primarily via blood. HCV may lead to chronic liver disease in over 50% of those virally infected. While immunoassay screening is often used as the first line of testing, HCV RNA testing assists in confirmation of active infection. Furthermore, in certain cases, molecular testing may be diagnostically indicated. For example, immunocompromised patients, such as those infected with Human immunodeficiency virus (HIV)-1, may not be seropositive for HCV but infected with the virus; thus immunoassays may result negative, while HCV RNA is detectable. Furthermore, because maternal antibodies can cross the placenta, qualitative molecular-based assays are more informative than serology in the detection of HCV RNA in the plasma or serum of infants born to HCV-infected mothers. RNA testing may also detect the virus earlier than immunoassays in any patient, as RNA is detectable as early as 1 week postexposure and several weeks prior to seroconversion, depending on the mode of transmission [26]. Such "qualitative" HCV RNA tests used for diagnosis commonly utilize conventional qPCR or transcription-mediated amplification, which, in contrast to traditional PCR, uses an isothermal reaction with RNA polymerase and reverse transcriptase (RT). Currently, available assays are detailed in Chapter 54, Infectious diseases; LODs may be as low as low as 5 IU/mL. Of note, RNAse activity of blood cells can lead to HCV RNA lability in whole blood, and rapid separation of serum is critical to avoid significant loss of HCV RNA and potential false-negative results [27].

For treatment, a quantitative viral load is typically assessed at baseline to predict the duration of treatment and the likelihood of sustained virologic response, which is defined as an undetectable viral load 24 weeks after completion of antiviral therapy. Commonly available formats for quantitative HCV RNA include both qPCR and bDNA assays. Quantitative assays also play a role in blood supply screening, which has significantly reduced the risk of HCV RNA transmission via transfusion reactions. While blood transfusion carried a high risk of 15%−20% for HCV infection prior to 1992, blood screening procedures have now reduced the risk dramatically to approximately 1 in 1 million transfusions [26].

HCV infection is genetically heterogeneous across individuals, with six major genotypes and over 100 subtypes identified. HCV genotyping assays serve to predict the course of infection and assist in determining treatment type and duration. Although the 5′ untranslated region is the most highly conserved portion of the HCV genome, it is frequently used as a target in all molecular HCV assays, including genotyping. Conversely, the NS5B nonstructural protein region of the HCV genome is the least well-conserved region, and more recently has been successfully used for viral genotyping. Common genotyping methodologies include reverse hybridization probe-based assays and direct DNA sequencing.

Human immunodeficiency virus type 1

HIV is an RNA lentivirus that is transmitted through the exchange of virus-containing fluids. Routes of infection include sexual contact, injection drug use, vertical transmission, and exposure to infected membranes or tissues. The type 1 virus (HIV-1) is the most common form worldwide and is the more virulent form than HIV-2. HIV-1 is characterized by extensive genetic heterogeneity, resulting in a variety of major phylogenetic groups (M, O, N) and M-group subtypes (A through K). The Centers for Disease Control and Prevention recommends that initial screening for HIV should be performed with an FDA-approved fourth-generation combination assay, which detects antibodies to both HIV-1 and HIV-2 as well as the HIV-1 p24 antigen [28]. However, RNA testing is indicated when serologic investigation is indeterminate or if an acute HIV infection is suspected. Virus may be detectable by the most sensitive molecular methods approximately 10 days after infection, possibly up to ten days prior to the fourth-generation immunoassays for the p24 antigen.

RNA quantification of HIV-1 viral load is primarily used for monitoring purposes in previously diagnosed HIV-1-infected individuals. Viral load testing can monitor response to therapy as well as predict time to progression to acquired immune deficiency syndrome (AIDS). Combinatorial antiretroviral (ARV) therapies seek to reduce viral loads below the LOD of currently available assays (20−50 copies/mL). Testing should be performed before therapy is initiated (baseline) and again at 2 and 8 weeks into treatment to assess response. Testing may then be performed every 3−6 months thereafter to evaluate continued therapeutic efficacy. If viral load exhibits a significant increase during therapy, HIV-1 resistance testing may be required (see below). While some individuals show persistently low but quantifiable levels of HIV RNA (<200 copies/mL but still detectable), these measurements are not necessarily indicative of virologic failure; the AIDS Clinical Trials Group defines virologic failure as a viral load of >200 copies/mL [29]. HIV-1 RNA quantification is typically achieved through qPCR or bDNA assays. FDA-approved assays for nucleic acid

TABLE 20.2 FDA-approved nucleoside reverse transcriptase inhibitors (NRTIs) for the treatment of HIV-1 and relevant resistance mutations in the HIV-1 reverse transcriptase (RT) gene.

NRTIs	Major HIV-1 RT mutations yielding resistance[a]	Notes
Zidovudine (ZDV/AZT) and stavudine (d4T)	T215Y/F, Q151M, T69ins	Thymidine analog mutations (TAMs) reduce susceptibility to these drugs
Lamivudine (3TC) and emtricitabine (FTC)	M184V/I	M184V is the most prevalent NRTI-associated mutation seen in HIV-1 treated patients
Tenofovir	K65R, T69ins	K65R confers significant cross-resistance to ABC, 3TC, and ddI
Abacavir (ABC)	K65R, L74V/I, Y115F, Q151M, T69ins	L74V may be present in combination with M184V
Didanosine (ddI)	K65R, L74V/I, Q151M, T69ins	The combination of d4T and ddI appears to select for TAMs more commonly than it selects for the L74V mutation

[a]*Mutations are described by a combination of letters and numbers, for example, M41L. The first letter (M) is the code for the amino acid in the wild-type virus. The number (41) identifies the position of the codon. The second letter (L) is the code for the new amino acid in the mutant sample. "ins" indicates an insertion.*
Source: Stanford University HIV Drug Resistance Database: http://hivdb.stanford.edu/DR/NRTIResiNote.html.

testing of HIV-1 are listed in Chapter 54, Infectious diseases, of this book. An additional technology, nucleic acid sequence-based amplification, is an isothermal method of RNA amplification that eliminates the need for thermocycling instruments and uses endpoint electrochemiluminescence or fluorescence detection for RNA quantification [30].

Viral drug resistance and genotyping

High rates of HIV replication and viral dependence on the error-prone HIV RT enzyme can lead to frequent introduction of viral mutations that often yield ARV drug resistance. To combat the rise of drug-resistant virus, combinatorial drug regimens are primarily used in the management of HIV, including inhibitors of CCR5-mediated viral entry, viral fusion, viral transcription, integration into the host cell genome, and protein assembly. However, viral resistance may occur in all of these ARV classes, thereby limiting the potential efficacy of treatment regimens. The likelihood of resistance is higher in certain ARV drug classes, such as nucleoside reverse transcriptase inhibitors (NRTIs)/nucleotide reverse transcriptase inhibitors and protease inhibitors. Thus there is immense clinical utility for antiviral resistance testing in HIV-1-infected patients, particularly in those who are on treatment regimens but not virally suppressed.

Traditional phenotypic assays measure the ability of the virus to grow among a variety of ARV drugs and concentrations. However, molecular assays are often cheaper, simpler, and provide a significantly quicker turnaround time; thus genotyping is the preferred method for evaluating resistance. The US Department of Health and Human Services recommends HIV drug resistance testing for individuals who have never before received therapy as well for those who are changing treatment regimens; however, this may not occur as standard of practice. Standard genotype-based drug-resistance testing involves detection of variants within the RT and PR genes. Assays assessing mutations in the integrase and gp41 (envelope) genes are also commercially available; however, the incidence of viral resistance in these genes is much lower. Automated sequencing methods are primarily used; upon isolation of the HIV-1 RNA from plasma, PCR amplification is performed followed by sequencing of the RT and PR genes. Most laboratories utilize commercial assays with corresponding software to assist in the interpretation of sequencing results. The International Antiviral Society—United States maintains and routinely updates lists of mutations associated with clinical resistance to HIV; the most recent report was published in 2019 [31]. As a representative example, Table 20.2 lists the NRTIs that are FDA-approved for the treatment of HIV-1 along with their majorly associated resistance mutations.

Assay approaches

Next-generation sequencing

While optimal for some applications, traditional sequencing methods are limited in throughput and are, therefore, not conducive to multigene panels. More recent technological advances in the early 2000s have led to the development of NGS, also known as MPS, which differs from

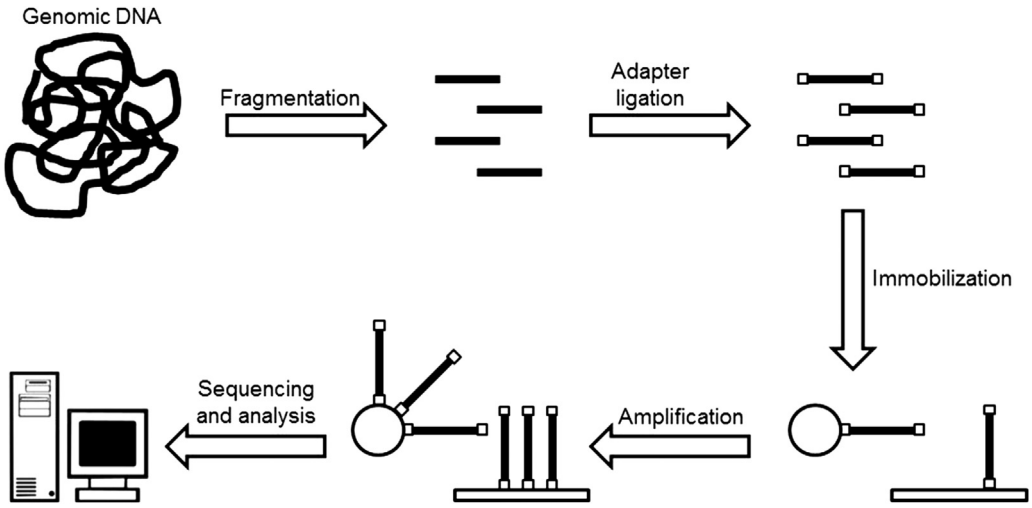


FIGURE 20.2 Next-generation sequencing general workflow. The general workflow of next-generation sequencing analysis is illustrated. Genomic DNA is first isolated, purified, and fragmented. Oligonucleotide adapters are attached to the ends of target fragments, and DNA is then immobilized to beads or sequencing slides. Fragments are then amplified clonally followed by parallel sequencing of the millions of clones, and resultant sequences are compared electronically against a reference genome. The whole genome may be sequenced, or smaller portions may be targeted, such as the exome or a preselected targeted gene panel.

traditional sequencing in that it uses parallel sequencing of millions of small DNA fragments to increase significantly throughput and speed of analysis. The technology has gained significant traction in the clinical laboratory and may be used in a number of ways. Notably, targeted NGS can provide both predictive and prognostic information in the routine care of patients with cancer. Further, NGS can also be used in the identification of germline variants associated with inherited diseases, identification of pathogens, analysis of circulating cfDNA, and gene discovery.

Though several technical platforms exist, NGS is generally a multiday process that first consists of a library preparation step in which the patient's DNA template is purified, amplified, fragmented, and then isolated onto solid surfaces such as slides or small beads (Fig. 20.2). Fragments are then amplified clonally followed by parallel sequencing of the millions of clones, and resultant sequences are compared electronically against a reference genome. The whole genome may be sequenced, or smaller portions may be targeted such as the exome or a preselected subset of genes ("hotspot" sequencing). Whole-genome sequence is typically reserved for gene discovery, including the characterization of unidentifiable diseases. Exome sequencing seeks to cover all coding regions of the genome. While the exome comprises only 1%–2% of the whole genome, it is estimated that it contains upward of 85% of disease-causing variants [32]. Exome sequencing may be indicated over more specifically targeted sequencing approaches when there are a large number of potential responsible genes in the patient's differential diagnosis or when obvious candidate genes have already been tested and were found to be normal. Targeted gene panels are selected when a particular disease is known or suspected, and the greater depth of coverage offered by these targeted panels increases the probability of detecting low-level heterogeneity or mosaicism. Notably, pipeline filters are often utilized to reveal only the clinically relevant or "actionable" variants.

Multiple commercial platforms are currently available for NGS analysis from a variety of vendors, including Illumina (HiSeq, MiSeq, and NextSeq series), Qiagen (GeneReader), and ThermoFisher Scientific (Ion Torrent), among others [33]. Considerations for choosing such a system include evaluation of read length, error rates, and run times. Notably, platforms vary with regard to gene coverage and may differ in terms of analytical sensitivity for specific variants. Thus inherent platform biases, combined with deviations in sample preparation and data analysis, can contribute to interlaboratory variability. Cost of testing and bioinformatics challenges remain hurdles in the wide implementation of NGS, though technological advances continue to decrease both.

Digital polymerase chain reaction

In the context of a dilute template DNA solution, such as in the case of cfDNA or fetal DNA in maternal plasma, the analytical sensitivity of conventional PCR is limited. Such limitations may be circumvented via the use of dPCR or single-molecule PCR. dPCR is performed through sample partitioning where template DNA is distributed broadly across thousands of reaction components. The physical dilution or partitioning of the sample occurs

to the extent that some compartments contain no template at all. Depending on the technical platform chosen, compartments may occur as microarrays, microfluidic disks, or small droplets of oil–water emulsions. Each partition behaves as an independent PCR reaction with the use of fluorescent probes to identify and quantify amplified target DNA. Data are analyzed in a binary way such that partitions containing product/fluorescent signal are considered positive ("1"), while partitions without detected signal are considered negative ("0"). Through the ratio of positive-to-negative reactions and Poisson statistics, very exact absolute template quantity can be determined. In this manner, copy number changes and minor allele fractions in a heavy background of wild-type DNA may be quantified. Applications are numerous in such areas as noninvasive prenatal diagnostics, liquid biopsies of solid tumor burden, and ultrasensitive viral load detection. While this method requires large-scale multiplexed PCR reactions, improvements in technologies such as microfluidics are making the technique more high throughput and suitable for routine use.

Universal biosensors for pathogen detection

The identification of unknown pathogens (bacteria, viruses, fungi, and protozoa) is challenging using conventional methods due to their often narrowly targeted nature as well as the long turnaround times associated with cultures. More recent technological advances have led to the development of universal pathogen detection platforms capable of identification and strain typing of a broad range of pathogens within the same assay. Abbott's IRIDICA universal biosensor (previously Ibis Biosciences) employs a series of broad-range primers for organism groupings, rather than particular species, to first PCR amplify species-specific genomic DNA. The broad-range approach of the primers allows for the inclusion of any species, even previously unknown organisms. Subsequent electrospray ionization mass spectrometry measures the mass-to-charge ratio (m/z) of the amplified product, which allows for the calculation of molecular mass for each amplicon. By comparing the resultant mass spectra to a database of known microorganisms, unknown pathogens can be matched and new, undescribed pathogens may be identified. The major advantage of this methodology is the ability to characterize multiple organisms simultaneously and without prior knowledge of microorganism identity. However, one caveat is that pathogen identification can only extend to what organisms are characterized in the reference library.

Regulatory considerations for molecular approaches

Clinically utilized molecular assays are subject to similar regulatory standards as other laboratory techniques, though they present their own set of challenges. The College of American Pathologists states that for molecular testing, nucleic acids must be extracted, isolated, and purified by commercially available or internally validated methods followed by measurement of nucleic acid quantity and purity. Internal quality control procedures should include the use of appropriate positive and negative controls in a similar matrix to patient samples. Laboratories must provide written guidelines for interpretation of results, which is often not straightforward with qualitative molecular assays or the detection of a novel genomic variant. Laboratory policies should be in place for the reporting of genetic findings unrelated to the intended purpose for clinical testing.

Regulations for NGS in particular pose unique challenges. "Positive controls" are typically chosen for one particular mutation or set of mutations and cannot encompass all possible mutations that could be recognized by the assay. In addition, Clinical Laboratory Improvement Amendments requirements for proficiency testing (PT) typically involve twice yearly assessments per assay. NGS assays may be defined in various manners, including a particular gene panel, exome analysis, or whole-genome analysis. The cost of running several PT samples per year may by any of these NGS assay definitions not be financially feasible for many laboratories. Alternatively, more affordable PT approaches that may become more prominent in the future include analytic/bioinformatics challenges in which electronic sequence information, rather than biological samples, is sent electronically to the laboratory. Such challenges may assess the laboratory's ability to accurately align, detect, and interpret the genetic information without having to use costly reagents and consumables. In general, developments in molecular technologies have initiated a shift from analyte-specific PT to methods-based PT due to the wide scope of genes now being assessed in clinical laboratories.

Limitations to molecular detection of allelic variants

Molecular diagnostic assays are accompanied by several technical and interpretative limitations that must be taken into consideration when used in the clinical laboratory. Sample quality comprises the first bottleneck of assay reliability. Nucleic acid extraction and sample preparation must be performed in sterilized environments to prevent contamination. For nucleic acids extracted from blood specimens, automated extraction methods are ideal for consistency of DNA sample preparation. For DNA obtained from solid tumors, the retrieval of DNA from formalin-fixed, paraffin-embedded tissue often leads to high levels of degradation and suboptimal sample quality. Resultant degraded samples may only allow for

amplification of small PCR amplicons (<200 base pairs). An additional barrier to biopsied tumor tissue is that isolated DNA will contain mixed populations of normal and cancerous cells. Thus assays must be sensitive enough to detect accurately very low levels of mutations in what may be a heavy background of normal sequences.

Aside from genome-wide approaches, a majority of molecular approaches are targeted to specific genes, gene regions, and known polymorphisms. It is imperative to recognize that such assays can detect only what they have been designed to recognize. Furthermore, bioinformatics challenges persist as advances in massive parallel sequencing assays yield more and more genetic information that requires large data storage capacity. In addition, questions still exist regarding how much information to transmit into the patient's medical record and how to best guide interpretation of that information. Finally, while costs per test are decreasing, many genetic tests remain expensive and may be deemed "investigational" by insurance companies and thus are not reimbursed.

Conclusions

This chapter has demonstrated the numerous applications of molecular techniques to the diagnosis and monitoring of patients in a variety of clinical disease states. Profound molecular breakthroughs in recent years have led to increased genotype−phenotype insights and their application in the clinical laboratory. However, with more knowledge come additional challenges in bioinformatics and test interpretation. Further collaborations among institutions and advancements in data analysis tools will soon allow for the routine testing for genetic mutations with broad, genome-wide approaches.

References

[1] E.R. Cantwell-Dorris, J.J. O'Leary, O.M. Sheils, BRAFV600E: implications for carcinogenesis and molecular therapy, Mol. Cancer Ther. 10 (2011) 385−394.

[2] H.H. Yoon, et al., KRAS codon 12 and 13 mutations in relation to disease-free survival in BRAF-wild-type stage III colon cancers from an adjuvant chemotherapy trial (N0147 alliance), Clin. Cancer Res. 20 (2014) 3033−3043.

[3] Competency Model for Pathologists. College of American Pathologists. <https://appsuite.cap.org/appsuite/learning/CompModel/Competency_Model.pdf>, 2019 (accessed 13.03.2020).

[4] A.C. Tsiatis, et al., Comparison of sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications, J. Mol. Diagn. 12 (2010) 425−432.

[5] A.G. Knudson, Two genetic hits (more or less) to cancer, Nat. Rev. Cancer 1 (2001) 157−162.

[6] H.T. Zheng, Z.H. Peng, S. Li, L. He, Loss of heterozygosity analyzed by single nucleotide polymorphism array in cancer, World J. Gastroenterol. 11 (2005) 6740−6744.

[7] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, *Nature* **409**, 2001, 860−921.

[8] P.E. Grundy, et al., Loss of heterozygosity for chromosomes 1p and 16q is an adverse prognostic factor in favorable-histology Wilms tumor: a report from the National Wilms Tumor Study Group, J. Clin. Oncol. 23 (2005) 7312−7321.

[9] S. Riethdorf, et al., Detection of circulating tumor cells in peripheral blood of patients with metastatic breast cancer: a validation study of the CellSearch system, Clin. Cancer Res. 13 (2007) 920−928.

[10] D. Mavroudis, Circulating cancer cells, Ann. Oncol. 21 (Suppl. 7) (2010) 95−100. vii.

[11] P. Mandel, P. Metais, Les acides nucleiques du plasma sanguin chez l'homme, C. R. Seances Soc. Biol. Ses. Fil. 142 (1948) 241−243.

[12] G.D. Sorenson, et al., Soluble normal and mutated DNA sequences from single-copy genes in human blood, Cancer Epidemiol. Biomarkers Prev. 3 (1994) 67−71.

[13] V. Vasioukhin, et al., Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia, Br. J. Haematol. 86 (1994) 774−779.

[14] Y.M. Lo, et al., Presence of fetal DNA in maternal plasma and serum, Lancet 350 (1997) 485−487.

[15] F.M. Lun, et al., Microfluidics digital PCR reveals a higher than expected fraction of fetal DNA in maternal plasma, Clin. Chem. 54 (2008) 1664−1672.

[16] Y.M. Lo, et al., Prenatal diagnosis of fetal RhD status by molecular analysis of maternal plasma, N. Engl. J. Med. 339 (1998) 1734−1738.

[17] Cell-free DNA. Screening for fetal aneuploidy. Committee Opinion No. 640. American College of Obstetricians and Gynecologists, *Obstet. Gynecol.* **126**, 2015, e31−e37.

[18] M.H. Chang, et al., Prevalence in the United States of selected candidate gene variants: Third National Health and Nutrition Examination Survey, 1991−1994, Am. J. Epidemiol. 169 (2009) 54−66.

[19] E.B. Spector, et al., Technical standards and guidelines: venous thromboembolism (Factor V Leiden and prothrombin 20210G >A testing): a disease-specific supplement to the standards and guidelines for clinical genetics laboratories, Genet. Med. 7 (2005) 444−453.

[20] E.M. Van Cott, B.L. Soderberg, M. Laposata, Activated protein C resistance, the factor V Leiden mutation, and a laboratory testing algorithm, Arch. Pathol. Lab. Med. 126 (2002) 577−582.

[21] P. Frosst, et al., A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase, Nat. Genet. 10 (1995) 111−113.

[22] C.J. Chu, et al., Hepatitis B virus genotypes in the United States: results of a nationwide study, Gastroenterology 125 (2003) 444−451.

[23] R.M. Madej, et al., International standards and reference materials for quantitative molecular infectious disease testing, J. Mol. Diagn. 12 (2010) 133−143.

[24] A. Valsamakis, Molecular testing in the diagnosis and management of chronic hepatitis B, Clin. Microbiol. Rev. 20 (2007) 426−439.

[25] A.S. Lok, B.J. McMahon, Chronic hepatitis B, Hepatology 45 (2007) 507−539.

[26] E. Gupta, M. Bajpai, A. Choudhary, Hepatitis C virus: screening, diagnosis, and interpretation of laboratory assays, Asian J. Transfus. Sci. 8 (2014) 19−25.

[27] D.H. Best, S.A. Dames, W. Wooderchak-Donahue, et al., Molecular pathology methods, in: D.G.B. Leonard (Ed.), Molecular Pathology in Clinical Practice, Springer, 2016, pp. 21–22.

[28] Laboratory Testing for the Diagnosis of HIV Infection: Updated Recommendations, Centers for Disease Control and Prevention, 2014.

[29] H.J. Ribaudo, et al., Racial differences in response to antiretroviral therapy for HIV infection: an AIDS clinical trials group (ACTG) study analysis, Clin. Infect. Dis. 57 (2013) 1607–1617.

[30] J.J. Weusten, W.M. Carpay, T.A. Oosterlaken, M.C. van Zuijlen, P.A. van de Wiel, Principles of quantitation of viral loads using nucleic acid sequence-based amplification in combination with homogeneous detection using molecular beacons, Nucleic Acids Res. 30 (2002) e26.

[31] A.M. Wensing, V. Calvez, F. Ceccherini-Silberstein, C. Charpentier, H.F. Günthard, R. Paredes, et al., Update of the drug resistance mutations in HIV-1, Top. Antivir. Med. 27 (3) (2019).

[32] H.L. Rehm, et al., ACMG clinical laboratory standards for next-generation sequencing, Genet. Med. 15 (2013) 733–747.

[33] S. Goodwin, J.D. McPherson, W.R. McCombie, Coming of age: ten years of next-generation sequencing technologies, Nat. Rev. Genet. 17 (6) (2016) 333–351.

Self-assessment questions

1. Which of the following techniques will detect copy-number neutral LOH?
   a. FISH
   b. karyotyping
   c. MSI
   d. pyrosequencing

2. The most frequent molecular genetic alteration observed in human cancers is:
   a. point mutations
   b. LOH
   c. translocations
   d. gene deletions

3. Mutations of prothrombin and Factor V are best described as:
   a. point mutations
   b. gene insertions
   c. gene deletions
   d. gene duplications

4. Molecular detection of hepatitis may be indicated in patients who are:
   a. anemic
   b. under the age of 18
   c. obese
   d. immunocompromised

5 Standardized HIV-1 drug resistance genotyping assays target the following genes:
   a. CCR5 and integrase
   b. PR and RT
   c. RT and integrase
   d. CCR5 and RT

Answers

1. c
2. b
3. a
4. d
5. b

Chapter 21

Applications of mass spectrometry in the clinical laboratory

Jennifer M. Colby[1] and Katie L. Thoren[2]

[1]*Department of Pathology, Microbiology, and Immunology, Vanderbilt University School of Medicine, Nashville, TN, United States,*

[2]*Department of Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, United States*

Learning objectives

After reviewing this chapter, the reader will be able to:

- Identify the advantages and disadvantages of mass spectrometry as a diagnostic tool.
- List three clinical small molecule and three clinical protein applications of mass spectrometry.
- Describe bottom-up and top-down approaches for clinical protein mass spectrometric assays.
- Describe the regulatory considerations associated with clinical mass spectrometric assays.

Mass spectrometry (MS), a technique in which the mass-to-charge ratio of a gas-phase ion is measured, has a long history in providing qualitative and quantitative measurements in analytical chemistry [1]. Most MS assays make use of isotopically labeled internal standards, which enable mass spectrometers to provide highly specific measurements. Mass spectrometers can be coupled to gas or liquid chromatographs to add orthogonal points of identification and further increase analytical and assay selectivity. The specificity of MS allows analytes to be measured accurately even in the presence of interfering substances (e.g., heterophilic or autoantibodies) or in the presence of structurally similar compounds that cannot be distinguished by immunoassay. Mass spectrometers are also capable of multiplexing, or measuring many analytes within a single analysis, and some parts of the testing process can be easily automated, both of which increase testing efficiency. A description of the analytical methodology is provided in Chapter 10, Mass spectrometry, of this book.

MS is an extremely versatile platform and offers many benefits as a measurement tool in laboratory medicine. In this chapter, we review the various applications of MS in the clinical lab as well as the challenges and the regulatory considerations associated with the implementation of this technique.

Mass spectrometric applications for measuring small molecules

Sample preparation for small molecule methods

Most MS methods require pretreatment of biological samples in order to introduce the sample into the mass spectrometer. For most methods currently deployed in clinical laboratories, sample preparation is used to concentrate an analyte of interest, dilute out interfering compounds, and/or change the composition of the sample to be more amenable for downstream chromatographic or MS analysis. Sample preparation for small molecule applications has been extensively reviewed elsewhere [2].

Gas chromatography-mass spectrometry (GC-MS) is generally limited to smaller, volatile compounds, and samples must be introduced in organic solvent. As a result, biological samples that will be analyzed by GC-MS are subjected to extraction, where the compound of interest is removed from the aqueous matrix and placed into an organic solution. A variety of techniques can accomplish this, including liquid−liquid extraction, solid phase extraction, and dilution. Extracted samples can be concentrated by evaporating some or all of the organic solvent. Extracts can also be chemically derivatized to increase the volatility of the analyte, improve the chromatographic peak shape, and increase the sensitivity of the method. Heat stable compounds that are generally nonpolar may not require derivatization prior to GC-MS analysis, but this process may be necessary for polar compounds. Derivatization reagents are often used for broad-spectrum drug screening methods, where compounds from a variety of chemical classes will be analyzed together, and for analysis of steroid hormones.

Depending on the type of chromatography being used, samples can be introduced into a liquid chromatography-

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00021-1

mass spectrometry (LC-MS) system in aqueous or organic solution. Liquid chromatography systems can be used with polar, nonpolar, and thermally labile compounds, so chemical derivatization is generally not necessary for LC-MS analysis. Still, it can be helpful to improve sensitivity for some analytes, like steroid hormones. Extraction is generally needed to remove proteins and lipids from serum and plasma, but is not absolutely required for matrices with lower protein content, like urine. In some cases, dilution of the specimen is all that is needed to achieve the desired sensitivity. In other cases, extraction followed by evaporation and reconstitution may be required to achieve optimal analytical sensitivity. The type of sample preparation that is selected will influence the degree of matrix interference, the frequency of source maintenance that is required, and ultimately the sensitivity of the method. These factors, as well as the cost of the preparation in technologist time, consumables, and solvents, can all influence the choice of sample preparation.

Qualitative and quantitative testing

Mass spectrometric testing used in diagnostics can be designed to operate qualitatively, for example, by identifying the presence but not the concentration of a compound of interest, or quantitatively, for example, by identifying both the presence and concentration of a compound of interest. Many factors can influence the choice to provide qualitative or quantitative results. Quantitative methods may require more extensive internal standardization and calibration and may be more expensive to develop, validate, and perform. In some clinical scenarios, like urine drug testing, quantitative results may not add value or may be misinterpreted, and qualitative results may be preferred. On the other hand, compounds that are expected to be present in a patient's sample, like steroid hormones and many other biomarkers, are generally measured quantitatively, and the numeric result is compared with a reference interval.

Toxicology

One of the first clinical applications of MS arose in part due to the 1981 accident on the U.S.S. Nimitz, in which cannabinoid metabolites were detected in urine samples collected from several crew members who were killed in the accident [3]. Intoxication was considered as a factor in the accident, but the presence of cannabinoids in the crew members' urine could not be used to determine unequivocally the timing of cannabis use. To prevent reoccurrence of this issue, the military enacted a zero-tolerance policy around drug use. As a consequence, increased demand for, and focus on, the accuracy of urine drug testing pushed laboratories to offer more specific confirmatory testing for presumptively positive immunoassay screens. Today, toxicological applications of MS have involved a variety of mass analyzers [quadrupole, ion trap, time of flight (TOF), and orbitrap] along with a variety of chromatography and ionization techniques [4]. Confirmation of drugs in biological fluids is an area where the specificity of MS is of particular importance, since the results can have medical and legal ramifications. Multiplexing of drug analytes is common in confirmatory mass spectrometric testing panels, which may contain a hundred or more drugs. Clinical testing is performed on a variety of biospecimens, including plasma, serum, urine, hair, and oral fluid. Other specimens may be used for specific purposes, for example, in screening newborns for in utero drug exposure. Confirmatory mass spectrometric results can be qualitative or quantitative. GC-MS and liquid chromatography tandem mass spectrometry (LC-MS/MS) are the most common techniques in laboratory medicine, with the use of LC-MS/MS growing quickly over the past decade. Most LC-MS/MS confirmation assays are targeted in nature, which increases the sensitivity of the analysis for compounds of interest. A more comprehensive review of toxicology-related applications can be found in Chapter 52, Toxicology and the clinical laboratory.

More recently, MS has been applied as a detection tool for novel psychoactive substances (aka "designer drugs" or NPS) [5]. NPS function similarly to known illicit drugs, but have slightly different structures; these differences lead to a legal grayzone, in which many of these compounds are often not explicitly illegal. Still, they can be dangerous, and pressure on the NPS market causes a continual shift in the exact drugs that are available. Because the targets change frequently and immunoassays can be costly and time-consuming to develop, many NPS can only be detected using MS. One of the most promising techniques used in this field is high-resolution MS, with a TOF or orbitrap mass analyzer coupled to upfront liquid chromatographic separation [5–7]. If data for all ions reaching the detector are collected, the data file can be analyzed retrospectively to detect specific compounds of interest. Though it is not as widely employed in laboratory medicine as the more traditional LC-MS/MS, high-resolution MS has much promise for emerging trends and public health threats.

Therapeutic drug monitoring

Therapeutic drug monitoring (TDM) is a type of personalized (or precision) medicine in which the blood concentration of a pharmaceutical agent is determined in an effort to ensure that the patient is experiencing the maximum benefit of the drug and to prevent toxicity (see Chapter 51: Therapeutic drug monitoring, for more detail).

While immunoassays are available for many drugs where TDM is suggested or required, there are a large number of drugs where commercial assays are unavailable, including antineoplastic agents like busulfan, anticonvulsants like rufinamide, antimicrobials like tazobactam and piperacillin, antifungals like itraconazole, and biologic drugs like infliximab. There are many other compounds, like antidepressants, that lack commercial immunoassays; in such cases, MS may be used to investigate the utility of TDM. One of the major advantages of mass spectrometric assays for the purpose of TDM is the specificity of the measurement in the presence of drug metabolites. Another major advantage is the ability to multiplex measurements of multiple drugs expected to be used to manage a patient's disease from a single injection. For many TDM applications, LC-MS/MS methods have been developed to quantitate drug concentrations from a blood matrix [4]. These methods are ideal for drugs like cyclosporine, in which immunoassays have variable cross-reactivity with active metabolites, potentially resulting in drastically different values [8]. In some cases, US Food and Drug Administration (FDA)-approved reagents for LC-MS/MS TDM assays are commercially available [9].

Endocrinology

After toxicology and TDM, measurement of hormones and related compounds usually makes up the next largest group of mass spectrometric methods in routine clinical laboratories. Given that many steroids share a common chemical backbone, immunoassays are prone to undesirable cross-reactivity [10,11]. In addition, some immunoassays lack the sensitivity to measure steroids in special populations. As a result of these factors, LC-MS/MS is often preferred for the measurement of testosterone in women, children, and hypogonadal men, and serves as a gold standard method for many other analytes (e.g., biogenic amines, 17-hydroxyprogesterone, aldosterone, and many more). From a laboratory perspective, one challenge associated with measuring endogenous compounds like those mentioned above is that matrix materials that are free of the analyte of interest must be acquired. These matrices can be produced from natural matrix, as is the case when serum is charcoal-stripped to remove steroid hormones from the matrix. Alternatively, surrogate matrices can be used, like bovine serum albumin in saline, which is created by the laboratory to mimic the protein content of the natural matrix. In many cases, these starting materials are expensive and may demonstrate significant lot-to-lot variability. Measurement of endogenous molecules adds an additional layer of complexity over exogenously derived molecules like drugs.

Newborn screening

Newborn screening targets and policies vary by state, and in some cases, testing is performed by dedicated state-run laboratories rather than hospital-based or associated laboratories. Due to the time-sensitive nature of newborn screening and the limited volume of sample that is collected for a large number of tests, the analytical sensitivity of MS is of particular value for this application [12,13]. Newborn screening is typically performed on dried blood spots rather than liquid samples, but similar to endocrinology testing, involves the detection of endogenously derived compounds present in abnormal amounts or patterns. Some compounds of interest like acylcarnitines are present as a number of closely related polymers, and specific tandem mass spectrometric data collection techniques called precursor ion scans are often used to detect these compounds. In many cases, newborn screening applications do not use gradient chromatography to separate analytes prior to introduction into the mass spectrometer. A variety of other biomolecules, including amino acids and organic acids are measured in newborn screening, and these analyses may be accomplished using traditional LC-MS/MS or GC-MS methods [12,13], refer to Chapter 48.

When an inborn error of metabolism is identified in newborn screening, treatment is often initiated, sometimes with pharmacologic agents, but very commonly with dietary restrictions. Mass spectrometric methods can also be used for routine monitoring of patients with known disease and can be helpful in ensuring that treatment regimens remain effective as the child grows. For patients with the most common amino acidopathy, phenylalanine hydroxylase deficiency, recommendations include keeping the diet low in phenylalanine and supplemented with tyrosine [14]. Similar to the TDM examples described above, patient compliance with dietary restrictions and supplementation can be monitored by measurement of amino acids in blood. Though many laboratories use liquid chromatography with UV detection to measure amino acids, LC-MS/MS offers a faster turnaround time, since analytes do not need to be completely resolved chromatographically [15]. This can be important for patients with inborn errors of metabolism like methylmalonic acidemia and maple syrup urine disease, who can become rapidly and acutely decompensated.

Trace elements

A variety of clinically relevant elements including lead, cobalt, chromium, zinc, and copper can be measured using MS. While many elements are tested infrequently, lead testing is commonly performed on children, because

lead has documented effects on development and the US Centers for Disease Control (CDC) has determined that there is no safe blood lead level in children [16]. Inductively coupled plasma mass spectrometry (ICP-MS) is the MS technique used for clinical metals testing. Similar to other MS applications, multiple metals can be measured simultaneously using ICP-MS, which can enable a laboratory to perform multiple analyses using a single workflow [17]. Samples for trace element testing are usually collected in certified metal-free tubes and ICP-MS testing can be performed in a room with limited access that is kept under positive-pressure to reduce environmental contamination. An overview of trace elements can be found in Chapter 46, Trace elements: functions and assessment of status through laboratory testing, of this book.

Mass spectrometry applications for measuring proteins

Although the mass spectrometric analysis of proteins and peptides is not very different from that of small molecules, the steps preceding analysis are more complex and currently represent a major bottleneck for translating these assays into the clinical laboratory. These steps are important, however, because biological matrices are complex mixtures of thousands of proteins, most of which are of low abundance. Strategies are needed to simplify the sample and enrich the protein or peptide of interest.

Techniques used in clinical proteomics

Bottom-up approach

A common approach for analyzing proteins by MS is bottom-up analysis. In this approach, proteins are enzymatically digested and peptides are used as a surrogate marker of the protein. Trypsin is commonly used for the digestion step because it reproducibly cleaves on the C-terminal (carboxy-terminal) side of arginine and lysine amino acid residues. The mass spectrometric analysis of peptides is similar to that of small molecules. Peptides are typically analyzed by LC-MS/MS. For purposes of quantitation, data are collected in a targeted manner. That is, the mass spectrometer only looks for certain peptides in an experiment called selected reaction monitoring or SRM (Fig. 21.1A). Here, the peptide precursor ion is selected and fragmented, and a specific product ion is detected. This precursor-product ion transition can be quantitated using internal standards and calibrators.

For the purposes of protein identification, data are collected in an untargeted manner (Fig. 21.1B). Because peptides fragment in a predictable manner (Fig. 21.2), MS/MS spectra obtained from a peptide can be searched against databases to identify proteins that are in the

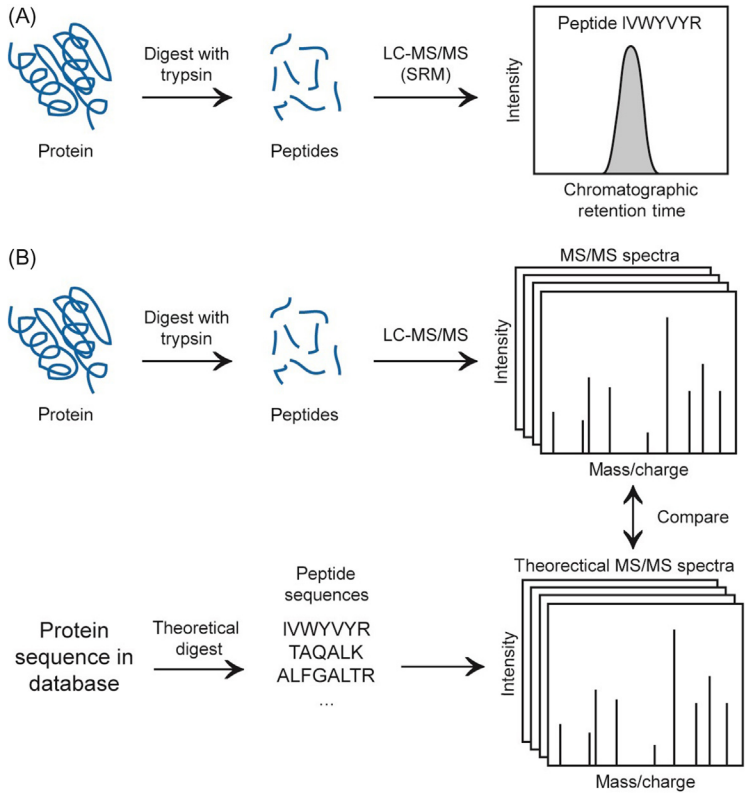


FIGURE 21.1 Targeted versus untargeted workflows for protein analysis. (A) In a targeted workflow, the mass spectrometer only looks for a specific peptide using selected reaction monitoring (SRM). (B) In an untargeted workflow, product ion spectra are generated from all peptides that enter the mass spectrometer. Because peptides fragment in a predictable manner, MS/MS spectra collected from a sample can be compared with theoretical spectra in order to identify peptides, and the corresponding proteins, in the sample.

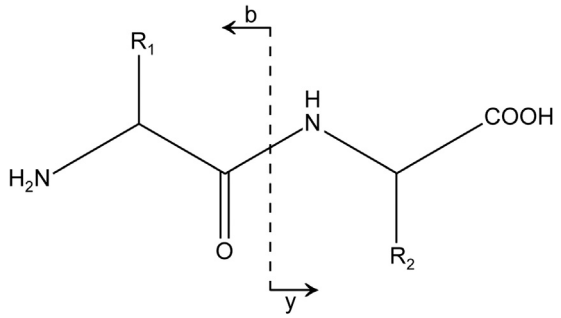


FIGURE 21.2 Fragmentation pattern of peptides. Under low-energy, collision-induced dissociation, peptides typically fragment at the peptide bond producing b- and y-ions, which contain the N-terminus and C-terminus, respectively.

sample or used for de novo protein sequencing. This approach is called "shotgun" proteomics and is used in discovery research to identify new biomarkers. This workflow is also being implemented for clinical purposes for the classification of amyloidosis, for example. The main advantages of using a bottom-up approach are that peptides ionize much more readily than large proteins because of their smaller size. The downside is that digestion introduces variability into the sample preparation step.

Top-down approach

In a top-down approach, intact proteins are measured. In some cases, proteins may be separated into smaller parts, either through reduction of disulfide bonds or dissociation of individual subunits, but the protein is not digested into peptides. The mass of the full-length protein or subunit, including posttranslational modifications, is measured. The advantage of the top-down approach is that the sample preparation step is simpler without proteolysis. In addition, different proteoforms and posttranslational modifications can be distinguished more easily by top-down analysis. The main disadvantage of this approach is that large proteins are harder to ionize. In addition, with electrospray ionization (ESI), large proteins may adopt many different charge states making mass spectra more complicated. Software programs are used to deconvolute the mass spectrum of multiply-charged species into a molecular mass (Fig. 21.3). Specific analytical platforms amenable to bottom-up or top-down approaches will be discussed later in this chapter.

Sample preparation

Many of the mass spectrometric methods employed to measure proteins rely on an upfront purification step to simplify the sample and enrich the protein or peptide of interest prior to analysis. Several strategies can be used for sample cleanup including solid-phase extraction, protein precipitation, ion-exchange isolation, liquid−liquid extraction, or depletion of high abundance proteins found in biological samples. A common strategy used in clinical assays is immunoaffinity enrichment. In this approach, antibodies that are immobilized to a solid support are used to purify the protein from the biological matrix. In bottom-up approaches, immunoaffinity enrichment can be performed at the protein or peptide level. That is, the protein can be purified first and then digested with trypsin. Alternatively, all proteins in a sample can be digested and then antipeptide antibodies are used to purify specific peptides. Purification of tryptic peptides from a digest is also known as the stable isotope standards and capture by antipeptide antibodies method [18].

Quantitation

As with small molecules, quantitation of proteins by MS is most often performed by the addition of internal standards. Ideally, isotopically labeled proteins would be used as internal standards and would be added early in the sample preparation process to control for the enzymatic digestion. However, labeled full-length proteins are expensive and usually not available. Instead, most bottom-up protein assays use stable isotope-labeled peptides that are added after digestion with trypsin. These peptides are chemically synthesized and include ^{13}C or ^{15}N heavy isotopes. Alternatively, some assays use cleavable, or extended, internal standards that include a proteolytic digestion site (Fig. 21.4). These labeled internal standards are added prior to trypsin digestion and provide a way to control for variation in the digestion step, though the digestion rate may not be representative of that of the full-length protein analyte.

Relative quantitation approaches may also be used. One example is spectral counting. When data are collected in an untargeted manner, proteins that are more abundant will produce more peptides and more MS/MS fragmentation spectra. Therefore the frequency of MS/MS events for a certain protein is used to determine its relative abundance in the sample. The advantage of relative quantitation is that it avoids the use of costly labeled internal standards and can be used when the identity of the protein(s) is not known. However, spectral counting is limited in terms of its accuracy, because the size of proteins and ionization efficiency of peptides can also affect the number of MS/MS events generated. Relative quantitation approaches are mostly used in research and discovery applications. However, one example of spectral counting being used in a clinical method is for the classification of amyloidosis, which is described in more detail below.

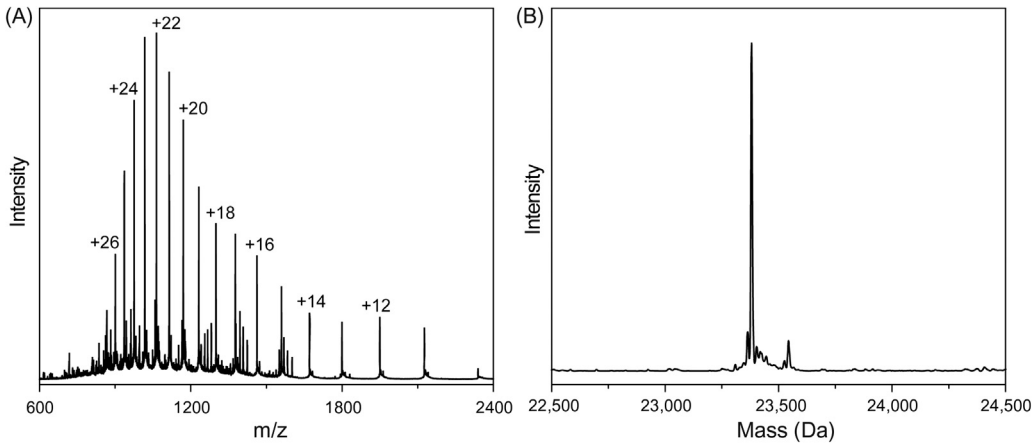


FIGURE 21.3 Electrospray ionization results in multiply charged protein ions. (A) When analyzed by liquid chromatography-electrospray ionization-Q-time-of-flight mass spectrometry, the intact kappa light chain of Daratumumab, an anti-CD38 therapeutic monoclonal antibody, appears as several multiply-charged ions. (B) Software is used to convert multiply charged ions into molecular mass, a process called deconvolution.

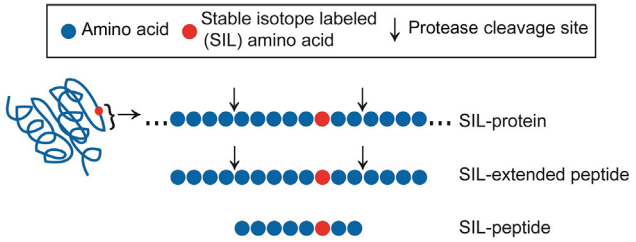


FIGURE 21.4 Internal standard strategies for protein quantification by mass spectrometry.

Instrumentation

Matrix-assisted laser desorption ionization (MALDI) and ESI are most often used for protein analysis because they are soft ionization techniques. This allows molecular ions to be obtained without extensive fragmentation during the ionization step, which is particularly important when ionizing large molecules. For protein analysis, triple quadrupole (QqQ), TOF, or orbitrap mass analyzers are often used. MALDI-TOF mass spectrometers are attractive for the analysis of intact proteins (i.e., top-down approach) because of their large mass range. ESI-QqQ is often used for targeted, quantitative analysis of peptides. ESI-Q-TOF and ESI-Q-Orbitrap systems are often used in the untargeted analysis of proteins, because they provide measurements with high resolution and high mass accuracy. Finally, nano- and microflow liquid chromatography systems are also common in protein analysis. These low-flow systems improve the sensitivity of the assay, because very small volumes of sample are delivered to the mass spectrometer, therefore reducing ion suppression. They also require smaller sample volumes. For more details on MS and ionization techniques, see Chapter 10, Mass spectrometry.

Examples of clinical protein mass spectrometric assays

A handful of MS-based protein assays are currently being used by clinical laboratories. Here, we provide examples that highlight the various approaches used and the reasons why MS is chosen over other techniques. Other examples are listed in Table 21.1.

Thyroglobulin

Thyroglobulin (Tg) is a 660-kDa protein that is produced by the follicular cells of the thyroid. Because thyroid tissue is the only source of Tg, Tg is monitored after total thyroidectomy or radioactive ablation to detect the persistence or recurrence of differentiated thyroid carcinoma. Most clinical laboratories use commercially available immunoassays to quantitate Tg. However, these assays are susceptible to interference by thyroglobulin antibodies (TgAb), which are present in 10%–30% of patients and typically result in falsely low Tg results.

MS has provided a solution to accurately measure Tg in the presence of TgAb. The TgAb interference is overcome because all proteins in the sample are digested with trypsin. Immunoaffinity enrichment is then used to purify Tg-specific peptides prior to their quantitation by LC-MS/MS. In patients without TgAb, MS measurements show excellent agreement to standard Tg immunoassays. In patients with TgAb present, MS-based Tg values tend to be higher than immunoassays [39].

Currently, a few major reference labs and academic centers offer Tg analysis by MS. All labs use the same general steps, that is, a bottom-up approach with immunoaffinity enrichment of Tg-specific peptides. However, there are differences in sample preparation, calibrators,

TABLE 21.1 Examples of clinical protein mass spectrometric assays.

Protein	What is measured?	References
Thyroglobulin	Tryptic peptides	[19–22]
IGF-1	Tryptic peptides	[23]
	Intact protein	[24,25]
IGF-2	Intact protein	[24]
Carbohydrate-deficient transferrin	Intact protein	[26]
Transthyretin	Intact protein	[27]
Hemoglobin	Intact alpha and beta subunits	[28]
	Tryptic peptides	[29,30]
Angiotensin I	Intact peptide	[31–33]
PTHrP	Intact peptide	[34]
Monoclonal immunoglobulins	Intact light chain	[35]
Infliximab	Tryptic peptides	[36]
Eculizumab	Intact light chain	[37]
Alpha-1 antitrypsin	Tryptic peptides	[38]

IGF, Insulin-like-growth-factor; *PTHrP,* parathyroid hormone-related peptide.

internal standards, LC-MS/MS platforms, and which Tg-specific peptide is used for quantitation. Despite these differences, the LC-MS/MS assays correlate well and efforts to harmonize the assays are ongoing [40].

Although MS is highly specific for the measurement of Tg and can overcome the aforementioned TgAb interference, the analytical sensitivity is not as good as currently available immunoassays. It is unclear if this difference in sensitivity is clinically relevant. Future work to evaluate the clinical sensitivity and specificity of these MS-based assays will be important in determining the impact of these tests on patient outcomes.

Insulin-like-growth-factor 1

Insulin-like-growth-factor 1 (IGF-1) is a 70 amino acid peptide hormone that is used to assess growth hormone (GH)-related disorders or to detect GH abuse for antidoping purposes. MS-based assays to measure IGF-1 were developed because immunoassays suffer from interference from IGF binding proteins and show poor agreement between platforms [23]. Several MS methods have been described. Currently, a small number of reference laboratories in the United States use a top-down approach to quantify intact IGF-1 using high-resolution mass spectrometers for analysis [24,25]. A bottom-up approach, which showed good agreement between different laboratories, is used to quantitate IGF-1 by World Anti-Doping Agency-accredited laboratories [23].

Monoclonal immunoglobulins (M-proteins)

Traditionally, electrophoretic methods (i.e., serum protein electrophoresis, capillary electrophoresis, and immunofixation) have been used to detect isotype M-proteins for the diagnosis and monitoring of multiple myeloma and other monoclonal gammopathies. However, with recent improvements in the treatment for multiple myeloma and the use of monoclonal antibody drugs, there is a growing need for more sensitive and specific laboratory tests. In addition, MS techniques, especially MALDI-TOF MS, could potentially offer higher throughput compared to electrophoretic methods.

Several MS-based methods have been developed to detect isotype M-proteins in serum and urine [35,41,42]. Currently, one reference lab in the United States is offering the MALDI-TOF assay as a replacement for serum immunofixation. Here, immunoaffinity enrichment is used to purify immunoglobulins from serum. In this assay, disulfide bonds are reduced in order to separate the heavy chains from the light chains, and the resulting mixture of intact heavy and light chains is analyzed by MS. The principle behind this approach is that each immunoglobulin has a unique amino acid sequence and, therefore, a unique mass. When overproduced, the M-protein will produce a “spike” in the mass spectrum that can be visualized above the polyclonal background (Fig. 21.5).

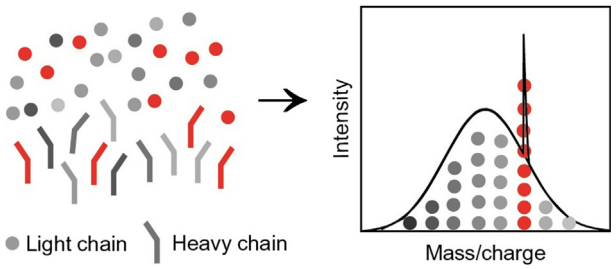


FIGURE 21.5 Monoclonal immunoglobulins can be visualized as a spike in the mass spectrum. In this assay, immunoglobulins are purified from serum or urine and are reduced to separate the light and heavy chains. When the sample is analyzed by matrix-assisted laser desorption ionization-time-of-flight mass spectrometry, the light chains generate most of the signal, since they ionize better than the heavy chains. The light chain from the monoclonal protein (*red*) can be detected as a spike above the polyclonal background (*gray*). The light chain has a specific amino acid sequence and, therefore, a specific mass and serves as a surrogate marker of the M-protein.

Classification of amyloidosis

Clinical management of amyloidosis is based on treating the underlying etiology and it is important to identify accurately the protein that is forming the amyloid deposits. An MS-based approach for classifying amyloidosis was described in 2009 [43] and is now being used by a few major reference laboratories in the United States. In this method, laser microdissection is used to cut out amyloid plaque from a tissue biopsy. Microdissected fragments are digested with trypsin, and the resulting peptides are analyzed by LC-MS/MS. The goal of this method is to identify the most abundant protein in the plaque; therefore the peptides are analyzed in an untargeted manner. Here, peptides are fragmented in the mass spectrometer and MS/MS spectra can then be searched against databases to identify the protein(s). Spectral counting is used to determine which protein(s) are most abundant in the sample. Laboratory professionals and pathologists review these results to determine if a known amyloidogenic protein is present. Results consist of an interpretive report, which indicates the presence and identity of the protein. This analysis is a significant improvement over immunohistochemistry techniques, which suffer from lack of sensitivity and specificity. In addition, the use of MS in this setting has identified previously unrecognized proteins in the formation of amyloid plaques [44].

Identification of microorganisms

The use of MALDI-TOF MS to identify microorganisms has drastically changed the clinical microbiology laboratory. This method is covered in more detail in Chapter 54, Infectious diseases, but we mention it here because the mass spectra generated in this assay are primarily comprised of highly abundant microbial proteins, such as ribosomal proteins [45]. Rather than trying to identify or quantitate the proteins in the sample, the mass spectrum is used as a "fingerprint" or "profile" of the microbe. The mass spectrum collected from a sample is searched against a database to identify organisms. "Fingerprint analysis" is also used in tissue imaging, which is discussed below in the "Future of clinical mass spectrometry" section.

Use of mass spectrometry in method standardization and harmonization

Standardization and harmonization of laboratory methods are important to ensure that different methods generate the same results when measuring the same sample. Harmonization refers to interlaboratory agreement or the "uniformity of test results when a reference method is not available" [46]. Standardization refers to agreement with the "true" value or the "uniformity of test results based on relation to a reference method" [46]. MS plays an important role in this area of laboratory medicine, because it directly detects molecules based on a well-defined physical property (i.e., mass). In addition, the use of isotopically labeled internal standards improves accuracy and precision and is a unique feature of the technique. Because of these advantages, isotope dilution MS is often used as a reference measurement method for a variety of analytes, including steroid hormones, drugs, metals/trace elements, and a few proteins. Please refer to the "Database of higher-order reference materials, measurement methods/procedures and services" [47], which is maintained by the Joint Committee for Traceability in Laboratory Medicine, for a full list of current reference measurement methods. Although no MS-based method has been fully standardized or harmonized according to recommended guidelines [48,49], progress has been made for some small molecules through the CDC's Hormone Standardization and Vitamin D Standardization Certification Program programs [50], for example.

Regulatory considerations for clinical mass spectrometry

The majority of MS-based assays currently deployed in clinical laboratories are considered laboratory developed tests (LDTs), which means the test is designed, manufactured, and used within a single laboratory [51]. The FDA has historically exercised discretion on regulating LDTs, with the regulatory oversight being provided by the laboratory's accrediting agency [51]. Recently, the FDA announced plans to regulate LDTs, and while the agency has not yet released final guidance, they have published a

TABLE 21.2 Guidelines pertaining to the design, development, validation, implementation, and use of mass spectrometry in laboratory medicine.

Title	Publisher	Publication year
Mass spectrometry in the clinical laboratory: general principles and guidance	CLSI	2007
Follow-up testing for metabolic diseases identified by expanded newborn screening using tandem mass spectrometry	AACC Academy (formerly NACB)	2008
Gas chromatography/mass spectrometry confirmation of drugs, 2nd edition	CLSI	2010
Liquid chromatography-mass spectrometry methods	CLSI	2014
Mass spectrometry for androgen and estrogen measurements in serum	CLSI	2015
Methods for the identification of cultured microorganisms using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry	CLSI	2017
Newborn screening by tandem mass spectrometry, 2nd edition	CLSI	2017
Using clinical laboratory tests to monitor drug therapy in pain management patients	AACC Academy (formerly NACB)	2017

AACC, American Association for Clinical Chemistry; *CLSI*, Clinical and Laboratory Standards Institute.

discussion paper summarizing the issues [52]. While the final guidance remains uncertain, it is clear that laboratories developing new LDTs will be subject to additional regulations. However, not all MS testing is developed by the laboratory. For example, there are several FDA-cleared/approved mass spectrometric devices for the purposes of neonatal screening, the identification of microorganisms, and the quantitation of exogenous and endogenous molecules, such as tacrolimus and vitamin D [53,54].

For laboratories utilizing LDTs, the increase in potential applications along with increased acceptance and awareness of the technique has spurned the development of several guidelines for best practices in MS method development and validation. Among other things, these guidelines describe experiments that should be conducted, provide suggestions for optimizing a method, and outline potential challenges that laboratories should consider in designing their methods. A full list of guidelines that may be useful for laboratories considering MS can be found in Table 21.2.

Challenges of implementing mass spectrometry in the clinical laboratory

Despite the advantages of MS over traditional techniques, there are many challenges associated with these assays. First, LDTs require a significant amount of time, money, and technical expertise to develop and validate [55]. There is a high upfront cost of instruments; technologists need to be well trained to prepare samples, operate the instruments and interpret data; and data analysis and reporting is time-consuming and may be manual. For protein analytes, the sample preparation step is one of the largest impediments to translating assays into the clinical laboratory. Regardless of the application, MS assays still require a substantial amount of manual input due to lack of automation. Because of these reasons, the use of MS, particularly for protein analytes, has largely been limited to reference laboratories and large academic centers.

Future of clinical mass spectrometry

Like many other techniques used in laboratory medicine, increasing automation is a goal for clinical MS and manufacturers are working to develop fully automated "black box" mass spectrometers. In this scenario, the entire MS analysis from sample preparation to result generation could be cleared by the FDA as a diagnostic test, which would make adoption of the technique much easier for laboratories with little experience. However, the mass spectrometer is programed only to run the analyses the manufacturer has validated; thus the laboratory may not have the option to develop and perform their own tests on the analyzer. Separately, software companies are working to improve the algorithms used to process MS data, a task that can be very time-consuming.

Mass spectrometers are also the primary tool for metabolomics, lipidomics, and proteomics research, in which hundreds of small molecules, lipids, or proteins/peptides

are measured [56]. The information from the MS analysis can be used to determine how disease-relevant biochemical pathways are functioning. Analysis of a pathway in patients with and without a target condition can identify candidate biomarkers that can be further validated. Highly accurate MS measurements can also be used to record the spatial distribution of various analytes within a tissue, thus creating an image based on the different masses collected [57]. MS is also being used in the operating room, including for the detection of surgical margins [58–60]. Though these analyses are currently only used for research purposes, they remain a point of interest for future applications of clinical MS. A variety of other potential applications have been reviewed more comprehensively in recent works [61].

Conclusion

Like any technique, MS has advantages and disadvantages. Developing and validating methods to the high standards required for diagnostic testing can be onerous and expensive. Data analysis is not always straightforward and may seem daunting for unexperienced laboratorians. However, MS enables laboratories to offer tests that they would otherwise have to send out and provides another tool with which the laboratory can contribute to patient care. When methods are validated with attention to best practices, MS can provide cost savings for laboratories, both directly in technologist time and indirectly in reduction of interferences from immunoassays, while also reducing the turnaround time. The arrival of automated mass spectrometers on the market, though still expensive, offer an opportunity for laboratories that do not have the technical capability or interest in LDT validation to still offer fast, selective, multiplexed analyses on their test menu.

References

[1] S. Borman, H. Russell, G. Siuzdak, A mass spec timeline: developing techniques to measure mass has been a Nobel pursuit, TCAW, 3, 2013, pp. 47–79.

[2] D.A. Wells, Sample Preparation for Applications. Principles and Applications of Clinical Mass Spectrometry [Internet], Elsevier, 2018, pp. 67–91. <https://linkinghub.elsevier.com/retrieve/pii/B9780128160633000037> (accessed 10.01.19).

[3] P.J. Jannetto, R.L. Fitzgerald, Effective use of mass spectrometry in the clinical laboratory, Clin. Chem. 62 (2016) 92–98.

[4] H.H. Maurer, Mass spectrometry for research and application in therapeutic drug monitoring or clinical and forensic toxicology, Ther. Drug Monit 40 (2018) 389–393.

[5] D. Pasin, A. Cawley, S. Bidny, S. Fu, Current applications of high-resolution mass spectrometry for the analysis of new psychoactive substances: a critical review, Anal. Bioanal. Chem. 409 (2017) 5821–5836.

[6] H.H. Maurer, M.R. Meyer, High-resolution mass spectrometry in toxicology: current status and future perspectives, Arch. Toxicol. 90 (2016) 2161–2172.

[7] A.H. Wu, R. Gerona, P. Armenian, D. French, M. Petrie, K.L. Lynch, Role of liquid chromatography–high-resolution mass spectrometry (LC-HR/MS) in clinical toxicology, Clin. Toxicol. 50 (2012) 733–742.

[8] A.J. McShane, D.R. Bunch, S. Wang, Therapeutic drug monitoring of immunosuppressants by liquid chromatography–mass spectrometry, Clin. Chim. Acta 454 (2016) 1–5.

[9] US Food and Drug Administration. Waters MassTrak Immunosuppressants Kit 510(k) Summary [Internet]. 2006. <https://www.accessdata.fda.gov/cdrh_docs/pdf6/K063868.pdf> (accessed 20.12.18).

[10] D. French, Advances in bioanalytical techniques to measure steroid hormones in serum, Bioanalysis 8 (2016) 1203–1219.

[11] S.S. Ketha, R.J. Singh, H. Ketha, Role of mass spectrometry in clinical endocrinology, Endocrinol. Metab. Clin. North Am. 46 (2017) 593–613.

[12] U. Garg, Y.V. Zhang, Mass spectrometry in clinical laboratory: applications in biomolecular analysis, in: U. Garg (Ed.), Clinical Applications of Mass Spectrometry in Biomolecular Analysis: Methods and Protocols [Internet], Springer New York, New York, NY, 2016, pp. 1–9. Available from: https://doi.org/10.1007/978-1-4939-3182-8_1.

[13] D. Ombrone, E. Giocaliere, G. Forni, S. Malvagia, G. la Marca, Expanded newborn screening by mass spectrometry: new tests, future perspectives: newborn screening: future directions, Mass Spectrom. Rev. 35 (2016) 71–84.

[14] D.S. Regier, C.L. Greene, Phenylalanine hydroxylase deficiency, in: M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J. Bean, K. Stephens, et al. (Eds.), GeneReviews((R)), University of Washington, Seattle, Seattle, WA, 1993.

[15] On behalf of the ACMG Laboratory Quality Assurance Committee, J.D. Sharer, et al., Laboratory analysis of amino acids, 2018 revision: a technical standard of the American College of Medical Genetics and Genomics (ACMG), Genet. Med. 20 (2018) 1499–1507.

[16] CDC Lead Poisoning Prevention Program: What do Parents Need to Know to Protect Their Children? [Internet]. United States Centers for Disease Control and Prevention. <https://www.cdc.gov/nceh/lead/acclpp/blood_lead_levels.htm>, 2017 (accessed 20.12. 18).

[17] M.A. Velez-Quinones, H. Xu, N. Vo, J.M. Gaitens, M.A. McDiarmid, M.R. Lewin-Smith, et al., Determination of thirteen trace and toxic elements in urine using inductively coupled mass spectrometry, Curr. Protoc. Toxicol. 78 (2018) e58.

[18] N.L. Anderson, N.G. Anderson, L.R. Haines, D.B. Hardie, R.W. Olafson, T.W. Pearson, Mass spectrometric quantitation of peptides and proteins using stable isotope standards and capture by anti-peptide antibodies (SISCAPA), J. Proteome Res. 3 (2004) 235–244.

[19] A.N. Hoofnagle, J.O. Becker, M.H. Wener, J.W. Heinecke, Quantification of thyroglobulin, a low-abundance serum protein, by immunoaffinity peptide enrichment and tandem mass spectrometry, Clin. Chem. 54 (2008) 1796–1804.

[20] M.M. Kushnir, A.L. Rockwood, W.L. Roberts, D. Abraham, A.N. Hoofnagle, A.W. Meikle, Measurement of thyroglobulin by liquid chromatography-tandem mass spectrometry in serum and plasma in the presence of antithyroglobulin autoantibodies, Clin. Chem. 59 (2013) 982–990.

[21] N.J. Clarke, Y. Zhang, R.E. Reitz, A novel mass spectrometry-based assay for the accurate measurement of thyroglobulin from patient samples containing antithyroglobulin autoantibodies, J. Investig. Med. 60 (2012) 1157–1163.

[22] B.C. Netzel, S.K.G. Grebe, A. Algeciras-Schimnich, Usefulness of a thyroglobulin liquid chromatography–tandem mass spectrometry assay for evaluation of suspected heterophile interference, Clin. Chem. 60 (2014) 1016–1018.

[23] H.D. Cox, F. Lopes, G.A. Woldemariam, J.O. Becker, M.C. Parkin, A. Thomas, et al., Interlaboratory agreement of insulin-like growth factor 1 concentrations measured by mass spectrometry, Clin. Chem. 60 (2014) 541–548.

[24] C. Bystrom, S. Sheng, K. Zhang, M. Caulfield, N.J. Clarke, R. Reitz, Clinical utility of insulin-like growth factor 1 and 2: determination by high resolution mass spectrometry, PLoS One 7 (2012). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3439428/> (accessed 5.01.19).

[25] H. Ketha, R.J. Singh, Quantitation of insulin-like growth factor 1 in serum by liquid chromatography high resolution accurate-mass massspectrometry, in: U. Garg (Ed.), Clinical Applications of Mass Spectrometry in Biomolecular Analysis: Methods and Protocols [Internet], Springer New York, New York, NY, 2016, pp. 131–137. Available from: https://doi.org/10.1007/978-1-4939-3182-8_15 (accessed 5.01.19).

[26] J.M. Lacey, H.R. Bergen, M.J. Magera, S. Naylor, J.F. O'Brien, Rapid determination of transferrin isoforms by immunoaffinity liquid chromatography and electrospray mass spectrometry, Clin. Chem. 47 (2001) 513–518.

[27] H.R. Bergen, S.R. Zeldenrust, M.L. Butz, D.S. Snow, P.J. Dyck, P.J.B. Dyck, et al., Identification of transthyretin variants by sequential proteomic and genomic analysis, Clin. Chem. 50 (2004) 1544–1552.

[28] I. Zanella-Cleon, P. Joly, M. Becchi, A. Francina, Phenotype determination of hemoglobinopathies by mass spectrometry, Clin. Biochem. 42 (2009) 1807–1817.

[29] F. Boemer, Y. Cornet, C. Libioulle, K. Segers, V. Bours, R. Schoos, 3-years experience review of neonatal screening for hemoglobin disorders using tandem mass spectrometry, Clin. Chim. Acta 412 (2011) 1476–1479.

[30] S.J. Moat, D. Rees, R.S. George, L. King, A. Dodd, A. Ifederu, et al., Newborn screening for sickle cell disorders using tandem mass spectrometry: three years' experience of using a protocol to detect only the disease states, Ann. Clin. Biochem. 54 (2017) 601–611.

[31] V.F. Fredline, E.M. Kovacs, P.J. Taylor, A.G. Johnson, Measurement of plasma renin activity with use of HPLC-electrospray-tandem mass spectrometry, Clin. Chem. 45 (1999) 659–664.

[32] C.E. Bystrom, W. Salameh, R. Reitz, N.J. Clarke, Plasma renin activity by LC-MS/MS: development of a prototypical clinical assay reveals a subpopulation of human plasma samples with substantial peptidase activity, Clin. Chem. 56 (2010) 1561–1569.

[33] J.G. Van Der Gugten, D.T. Holmes, Quantitation of plasma renin activity in plasma using liquid chromatography-tandem mass spectrometry (LC-MS/MS), Methods Mol. Biol. 1378 (2016) 243–253.

[34] M.M. Kushnir, A.L. Rockwood, F.G. Strathmann, E.L. Frank, J.A. Straseski, A.W. Meikle, LC-MS/MS measurement of parathyroid hormone-related peptide, Clin. Chem. 62 (2016) 218–226.

[35] J.R. Mills, M.C. Kohlhagen, S. Dasari, P.M. Vanderboom, R.A. Kyle, J.A. Katzmann, et al., Comprehensive assessment of M-proteins using nanobody enrichment coupled to MALDI-TOF mass spectrometry, Clin. Chem. 62 (2016) 1334–1344.

[36] M.A.V. Willrich, D.L. Murray, D.R. Barnidge, P.M. Ladwig, M. R. Snyder, Quantitation of infliximab using clonotypic peptides and selective reaction monitoring by LC–MS/MS, Int. Immunopharmacol 28 (2015) 513–520.

[37] P.M. Ladwig, D.R. Barnidge, M.A.V. Willrich, Quantification of the IgG2/4 kappa monoclonal therapeutic eculizumab from serum using isotype specific affinity purification and microflow LC-ESI-Q-TOF mass spectrometry, J. Am. Soc. Mass Spectrom. 28 (2017) 811–817.

[38] Y. Chen, M.R. Snyder, Y. Zhu, L.J. Tostrud, L.M. Benson, J.A. Katzmann, et al., Simultaneous phenotyping and quantification of α-1-antitrypsin by liquid chromatography-tandem mass spectrometry, Clin. Chem. 57 (2011) 1161–1168.

[39] A.N. Hoofnagle, M.Y. Roth, Clinical review: improving the measurement of serum thyroglobulin with mass spectrometry, J. Clin. Endocrinol. Metab. 98 (2013) 1343–1352.

[40] B.C. Netzel, R.P. Grant, A.N. Hoofnagle, A.L. Rockwood, C.M. Shuford, S.K.G. Grebe, First steps toward harmonization of LC-MS/MS thyroglobulin assays, Clin. Chem. 62 (2016) 297–299.

[41] D.R. Barnidge, S. Dasari, C.M. Botz, D.H. Murray, M.R. Snyder, J.A. Katzmann, et al., Using mass spectrometry to monitor monoclonal immunoglobulins in patients with a monoclonal gammopathy, J. Proteome Res. 13 (2014) 1419–1427.

[42] D.R. Barnidge, R.C. Tschumper, J.D. Theis, J.D. Theis, M.R. Snyder, D.F. Jelinek, et al., Monitoring M-proteins in patients with multiple myeloma using heavy-chain variable region clonotypic peptides and LC–MS/MS, J. Proteome Res. 13 (2014) 1905–1910.

[43] J.A. Vrana, J.D. Gamez, B.J. Madden, J.D. Theis, H.R. Bergen, A. Dogan, Classification of amyloidosis by laser microdissection and mass spectrometry–based proteomic analysis in clinical biopsy specimens, Blood 114 (2009) 4957–4959.

[44] A. Dogan, Amyloidosis: insights from proteomics, Annu. Rev. Pathol. 12 (2017) 277–304.

[45] R. Patel, MALDI-TOF MS for the diagnosis of infectious diseases, Clin. Chem. 61 (2015) 100–111.

[46] Harmonization of clinical laboratory test results [Internet]. American Association for Clinical Chemistry. <https://www.harmonization.net/media/1087/aacc_harmonization_position_statement_2013.pdf>, 2013 (accessed 20.12.18).

[47] Database of higher-order reference materials, measurement methods/procedures and services [Internet]. <https://www.bipm.org/jctlm/> (accessed 10.01.19).

[48] R.F. Greaves, A guide to harmonisation and standardisation of measurands determined by liquid chromatography – tandem mass

spectrometry in routine clinical biochemistry, Clin. Biochem. Rev. 33 (2012) 123–132.

[49] A.L. Rockwood, M.S. Lowenthal, C. Bystrom, Harmonization of liquid chromatography–tandem mass spectrometry protein assays, Clin. Lab. Med. 38 (2018) 499–513.

[50] HoSt/VDSCP: Standardization of measurement procedures [Internet]. <https://www.cdc.gov/labstandards/hs_standardization.html>, 2018 (accessed 10.01.19).

[51] Laboratory developed tests [Internet]. United States Food and Drug Administration. <https://www.fda.gov/medicaldevices/productsandmedicalprocedures/invitrodiagnostics/laboratorydevelopedtests/default.htm>, 2018. (accessed 20.12.18).

[52] Discussion paper on laboratory developed tests (LDTs) [Internet]. United States Food and Drug Administration. <https://www.fda.gov/downloads/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/LaboratoryDevelopedTests/UCM536965.pdf>, 2017 (accessed 20.12.18).

[53] Sciex topaz total 25-hydroxyvitamin D mass spectrometry test system 510(k) summary [Internet]. United States Food and Drug Administration. <https://www.accessdata.fda.gov/cdrh_docs/pdf17/DEN170019.pdf>, 2017 (accessed 20.12.18).

[54] J.T. Lathrop, D.A. Jeffery, Y.R. Shea, P.F. Scholl, M.M. Chan, US Food and Drug Administration perspectives on clinical mass spectrometry, Clin. Chem. 62 (2016) 41–47.

[55] B.A. Rappold, Special considerations for liquid chromatography–tandem mass spectrometry method development, Clin. Lab. Med. 38 (2018) 539–551.

[56] A.D. Kennedy, B.M. Wittmann, A.M. Evans, L.A.D. Miller, D.R. Toal, S. Lonergan, et al., Metabolomics in the clinic: a review of the shared and unique features of untargeted metabolomics for clinical research and clinical testing, J. Mass Spectrom. 53 (2018) 1143–1154.

[57] M.M. Gessel, J.L. Norris, R.M. Caprioli, MALDI imaging mass spectrometry: spatial molecular analysis to enable a new age of discovery, J. Proteom. 107 (2014) 71–82.

[58] E.R. St John, J. Balog, J.S. McKenzie, M. Rossi, A. Covington, L. Muirhead, et al., Rapid evaporative ionisation mass spectrometry of electrosurgical vapours for the identification of breast pathology: towards an intelligent knife for breast cancer surgery, Breast Cancer Res. [Internet] 19 (2017). <http://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-017-0845-2> (accessed 10.01.19).

[59] C.R. Ferreira, K.E. Yannell, A.K. Jarmusch, V. Pirro, Z. Ouyang, R.G. Cooks, Ambient ionization mass spectrometry for point-of-care diagnostics and other clinical measurements, Clin. Chem. 62 (2016) 99–110.

[60] D.L. Phelps, J. Balog, L.F. Gildea, Z. Bodai, A. Savage, M.A. El-Bahrawy, et al., The surgical intelligent knife distinguishes normal, borderline and malignant gynaecological tissues using rapid evaporative ionisation mass spectrometry (REIMS), Br. J. Cancer 118 (2018) 1349–1358.

[61] A. Rockwood, The future of clinical mass spectrometry, Clin. Lab. N. 3 (2015).

Self-assessment questions

1. Which of the following is an advantage of mass spectrometry?
   a. Low sensitivity
   b. Low specificity
   c. Ability to multiplex
   d. Minor regulatory burden
2. What challenge is more often associated with measuring testosterone than with measuring cyclosporine?
   a. Sourcing testosterone-free matrix materials
   b. Sourcing reference materials for testosterone
   c. Detecting signal generated by testosterone
   d. Standardizing the assay against a reference material
3. Of the following applications, which would be most appropriate for analysis by ICP-MS?
   a. Illicit drugs
   b. Testosterone
   c. Phenylalanine
   d. Copper
4. Why does mass spectrometry help enable harmonization?
   a. Because everyone uses the same method to measure the analyte.
   b. Because everyone directly measures the analyte's signal.
   c. Because the ionization process is consistent across different mass spectrometers.
   d. Because many analytes can be detected in a single injection.
5. Which ionization technique is often used for the analysis of intact proteins?
   a. MALDI
   b. Electron ionization
   c. Fast-atom bombardment
   d. Atmospheric pressure chemical ionization
6. Which endopeptidase is commonly used in bottom-up analysis of proteins?
   a. Chymotrypsin
   b. Pepsin
   c. Trypsin
   d. Elastase
7. In low-energy, collision-induced dissociation, peptides typically produce which type of fragment ions?
   a. a- and b-ions
   b. b- and y-ions
   c. x- and y-ions
   d. Cations

Answers

1. c
2. a
3. d
4. b
5. a
6. c
7. b

Chapter 22

Proteins: analysis and interpretation in serum, urine, and cerebrospinal fluid

Joesph R. Wiencek[1], Show-Hong Duh[2], and Robert H. Christenson[2,3]

[1]Department of Pathology, University of Virginia School of Medicine, Charlottesville, VA, United States, [2]Department of Pathology, University of Maryland School of Medicine, Baltimore, MD, United States, [3]Department of Medical and Research Technology, University of Maryland School of Medicine, Baltimore, MD, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- List the clinical indications for requesting or following specific protein measurements or protein electrophoresis.
- Discuss and sketch a conventional protein electrophoresis pattern, label each fraction, and list two proteins that comprise each region.
- Describe common analytical methods and biochemical techniques used in protein analysis.
- List and classify potential endogenous and exogenous interferences in protein analysis and interpretation.
- Explain the physiology of serum free light chains and list three reasons why testing is clinically important.
- Explain clinical and laboratory features for the following conditions: multiple myeloma, Waldenström macroglobulinemia, amyloid light-chain amyloidosis, cryoglobulinemia, Bence Jones proteinuria, and monoclonal gammopathy of undetermined significance.
- Describe a basic strategy for evaluating pathological and nonpathological protein electrophoresis patterns and problem-solving when potential interferences are suspected.t

Introduction

Analysis and interpretation of proteins in serum, urine, and cerebrospinal fluid (CSF) are essential components for care of patients. It is vital that the laboratory provides informed consultation concerning technical performance, reference intervals, and serum protein interpretations in disease and potential interferences such as the appearance of a therapeutic monoclonal antibody (t-mAb) on protein electrophoresis. Some important indications for clinicians requesting specific protein measurements and protein electrophoresis are the following:

1. workup and monitoring of patients for monoclonal gammopathy;
2. as part of the workup of patients with suspected or diagnosed multiple sclerosis (MS);
3. monitoring of patients to evaluate response to therapy and prognosis;
4. unexplained renal insufficiency or proteinuria;
5. hypogammaglobulinemia;
6. peripheral neuropathy;
7. clinical history of amyloid light-chain amyloidosis; and
8. further evaluation of patients having unexpected findings, for example, bone lesions on X-ray performed for unrelated indications.

Fig. 22.1 displays the major regions of the typical serum protein electrophoresis (SPEP or SPE) pattern for a normal individual. Table 22.1 lists various proteins and the corresponding SPEP region where they are located. Each region will be discussed in the order of their migration pattern on SPEP, from the anode to the cathode. Reference to some proteins will be made in the context of the "acute-phase response" (APR). This term refers to the dynamic changes in these protein concentrations in response to physiologic stresses such as inflammation, trauma, myocardial infarction, or cancer.

Prealbumin region

Prealbumin or "transthyretin" (TTR) serves as a transport protein for thyroid hormone thyroxine (T_4) and retinol binding protein 4 (RBP4) in serum and CSF. The name TTR originates from the combination of the root words *trans*ports *thy*roxine and *retin*ol, but occasionally is referred to as prealbumin, because it migrates faster than albumin (ALB) on protein electrophoresis. TTR is synthesized by both the liver and choroid plexus and appears more prominently on CSF

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00022-3

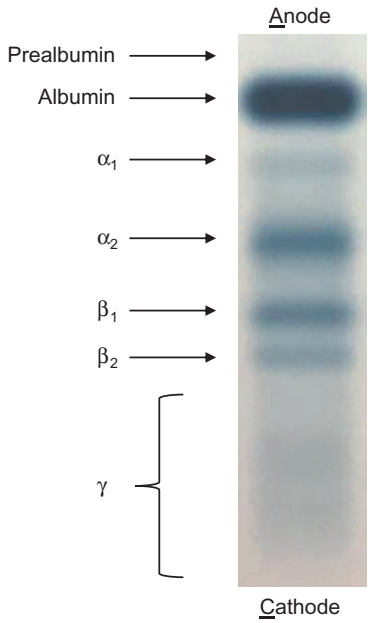


FIGURE 22.1 Normal serum protein electrophoresis pattern performed in an agarose matrix. The six typical fractions separated are indicated (albumin, α_1, α_2, β_1, β_2, γ).

protein electrophoresis than on SPEP except in patients receiving certain medications (e.g., salicylates). Since the concentration of TTR decreases with physiological stress, it is often referred to as a negative APR protein. Table 22.1 lists the characteristics of various proteins, including TTR, such as its relatively short half-life of 48 hours. The amino acid sequence of TTR contains a high proportion of essential amino acids, and this protein is rapidly degraded during nutritional deficit and serves as a reservoir for these amino acids. It is important for clinical laboratorians to know that prealbumin and TTR may be used interchangeably.

Another protein in the prealbumin region is RBP4. This protein functions in the transport of the all-trans form of the fat-soluble vitamin retinol (vitamin A). RBP4 has a short half-life and circulates as a dimer bound to TTR, which stabilizes the interaction between RBP4 and vitamin A and helps conserve this relatively small protein from renal excretion. The liver synthesizes RBP4 in response to retinol binding requirements. RBP4 is cleared via renal excretion and, therefore, renal insufficiency leads to increased concentrations. Recent studies have demonstrated that increased concentrations of RBP4 may be an additional risk factor for type 2 diabetes and cardiovascular diseases [1–3]. Decreased RBP4 concentrations occur with liver failure and in protein malnutrition.

Albumin region

ALB is the most abundant protein in plasma, normally accounting for approximately 60%–65% of total protein and 80% of oncotic pressure. At physiologic pH, ALB is negatively charged with many surface groups and contains both hydrophilic and hydrophobic binding sites. This allows ALB to function as a carrier protein for various substances, including divalent cations, bilirubin, free fatty acids, and hormones such as thyroxine and estrogen, along with drugs such as penicillin, phenytoin, and warfarin. ALB also functions as an amino acid reservoir, entering cells by pinocytosis where ALB is available for degradation into constituent amino acids for the synthesis of cellular proteins. ALB's low isoelectric point (pI) confers an intense negative charge at the physiologic pH of 7.4 and migrates anodally on conventional SPEP (see Fig. 22.1). The ALB band can be used as the reference point in SPEP pattern interpretation.

The liver synthesizes ALB, and reduction of over 90% of the liver's functionality is necessary to decrease substantially the serum ALB concentrations. For this reason, acute hepatic disease is rarely a cause of decreased ALB. In inflammatory disease, serum ALB decreases, because synthesis is inhibited by the cytokines interleukin-1 (IL-1) and interleukin-6 (IL-6), and/or tumor necrosis factor-α (TNF-α). Given that there is an acute decrease in ALB concentrations in the inflammatory state, ALB is often referred to as a negative APR. The primary cause of hyperalbuminemia is dehydration; false hyperalbuminemia may also result from keeping a tourniquet on for too long during phlebotomy.

The interpretation of ALB on SPEP or urine protein electrophoresis (UPEP) is an important consideration in evaluating the renal status of a patient. ALB normally is the most predominant protein excreted in urine, because it is the most abundant plasma protein and also because it is not entirely reabsorbed by the renal tubules. Less abundant and smaller proteins may pass through the glomeruli, but most are almost entirely reabsorbed by the proximal tubules.

The size threshold for passage through the glomerulus is approximately 50 kilodaltons (kDa), only slightly less than ALB's molecular weight (67 kDa). ALB's strong negative charge at physiologic pH of 7.4 is important in minimizing its excretion through the kidney, because the glomerular basement membrane is also negatively charged, resulting in the electrical repulsion of ALB.

In a healthy state, any ALB that passes through the glomerulus is almost totally reabsorbed by the proximal tubule by active transport through receptor-mediated endocytosis. When there is an injury to the glomerular basement membrane and/or the renal tubules, substantial amounts of ALB and other proteins that are normally reabsorbed are excreted into the urine and can be visualized on UPEP. In nephrotic syndrome, where the amount of protein (mainly ALB) passing through the glomerulus increases to exceed the maximum capacity of tubular

TABLE 22.1 Characteristics of major proteins separated by serum protein electrophoresis.

Location on gel electrophoresis	Protein	Half-life	Molecular weight, g/mol	Isoelectric point	Normal reference interval (mg/dL)	Acute-phase protein	Comments
Anode (charge +)							
Prealbumin	TTR	48 h	54,000	4.7	20–40	Yes, ↓	Transport protein, nutrition marker
	RBP4	12 h	21,000		2.4–6.7		Nutrition marker, binds to TTR
Albumin	ALB	19 d	66,000	~5.0	3500–5200	Yes, ↓	Transport protein, maintains oncotic pressure and amino acid reserve
	A1AT	4 d	54,000	4.8	90–200	Yes, ↑	Genetic deficiency can result in emphysema or cirrhosis
α_1	AGP (orosomucoid)	5 d	40,000	3.5	50–120	Yes, ↑	Binds hormone and cationic drugs (function not fully known)
	Hp	2 d	85,000	4.1	30–200	Yes, ↑	Binds to hemoglobin and decreased in hemolysis
α_2	A2M	5 d	800,000	5.4	130–300	Yes, ↓	Proteolytic enzyme inhibitor, increased in nephrotic syndrome
	Cp	4.5 d	160,000	4.4	20–60	Yes, ↑	Antioxidant, contains copper, and decreased in Wilson's disease
β_1	Tf	7 d	77,000	5.7	200–360	Yes, ↓	Iron transport protein; increased in iron deficiency anemia
β_2	C3		180,000		90–180		Complement factor
	B2M	2 h	11,800		< 0.27		Assess renal tubular function and tumor marker
	IgG	24 d	160,000	6–7.3	700–1600		Long-term defense and most abundant serum immunoglobulin
γ	IgA	6 d	170,000		70–400		Active in secretions and increases in B-cell tumors
	IgM	5 d	900,000		40–230		First immunoglobulin to increase (respond) after infection
	CRP	18 h	120,000	6.3	< 0.5	Yes, ↑	Robust and nonspecific APR protein
Cathode (charge −)							

A1AT, α_1-Antitrypsin; *A2M*,α_2-macroglobulin; *AGP*, α_1- acid glycoprotein; *ALB*, albumin; *B2M*, β_2 microglobulin; *C3*, complement 3; *Cp*, ceruloplamin; *CRP*, C-reactive protein; *Hp*, haptoglobin; *IgA*, immunoglobulin A; *IgG*, immunoglobulin G; *IgM*, immunoglobulin M; *RBP4*, retinol binding protein 4; *Tf*, transferrin; *TTR*, transthyretin; α_2, alpha 2; β_1, beta 1; β_2, beta 2; γ, gamma.

resorption, damaged glomeruli will allow 3 g or more of protein to be excreted in urine over a 24-hour period [4]. The loss of ALB leads to reduction in oncotic pressure and excess fluid buildup in tissues. This may lead to the synthesis of large proteins such as α_2-macroglobulin (A2M) and apolipoproteins as a compensatory response to maintain oncotic pressure. This results in the increase in the peak size of the α_2 region of the SPEP.

Two additional patterns in the ALB region to be cognizant of when interpreting protein electrophoresis patterns are from patients with analbuminemia or bisalbuminemia. Analbuminemia (OMIM 103600) is a rare autosomal recessive disorder characterized by low concentrations of ALB and the absence of an ALB band on SPEP. The disorder is accompanied by mild symptoms of edema and fatigue. Bisalbuminemia or alloalbuminemia appears as a doublet in the ALB region and may be inherited or acquired (Fig. 22.2). The inheritance pattern of bisalbuminemia is through an autosomal codominant pattern and is considered relatively benign. Cases of the acquired form are described in patients receiving high doses of β-lactam antibiotics or those with pancreatic disease.

α_1 Region

The major protein in this region is α_1-Antitrypsin (A1AT). A1AT functions as a "broad spectrum" inhibitor for trypsin, elastase, and several other proteases released by neutrophils. A1AT can also inhibit members of the clotting cascade. Synthesized in the liver, A1AT rises rapidly in response to the increased concentrations of IL-6 and other proinflammatory cytokines. A1AT is a robust positive APR protein that may function to attenuate proteases and other substances that are released during inflammation. Catabolism of A1AT occurs via receptors located on hepatic parenchymal cells, which bind A1AT-protease complexes.

A decrease in α_1 globulins may be indicative of A1AT deficiency. The gene coding for A1AT (termed "Pi") is polymorphic, and there are over 75 genetic variants or isotypes. Although many of these variants are associated with reduced concentrations of A1AT, the most clinically important is the Z allele denoted as PiZ. The homozygous PiZ genotype (e.g., PiZZ) occurs in 1 in 6700 live births of individuals having northern European ancestry. Approximately 90% of these individuals develop emphysema between the second and fourth decades of life. However, many individuals go undiagnosed, especially in patients with chronic obstructive pulmonary disease. Individuals who are heterozygous or homozygous for the PiZ allele have an increased incidence of cirrhosis and hepatocellular carcinoma.

Detection of PiZ and other variants of A1AT can be accomplished by isoelectric-focusing electrophoresis (IEF); this is considered the biochemical gold standard test for establishing the diagnosis of A1AT deficiency. However, caution must be taken, because if plasma is not separated expeditiously from cellular material, leukocytes may release elastase and other proteases that can bind with A1AT and change electrophoretic mobility. This in vitro binding may also affect immunochemical measurement. In severe cases of A1AT deficiency, there may be an absence of the α_1 region on SPEP. Additional confirmatory tests may include targeted genotyping, protease inhibitor IEF, mass spectrometry, or gene sequencing. Treatments include transfusion of purified A1AT, aerosols, liver transplantation, and, perhaps in the future, gene therapy.

α_1-Acid glycoprotein (AGP), also termed "orosomucoid," is synthesized by the liver and in granulocytes and monocytes. Catabolism of AGP occurs via a receptor-mediated mechanism involving hepatic parenchymal cells. The physiologic function of AGP is incompletely understood, but the protein is known to bind progesterone and other hormones, as well as drugs, such as lidocaine, propranolol, and phenobarbital. AGP is a positive APR protein and has been used for monitoring ulcerative colitis.

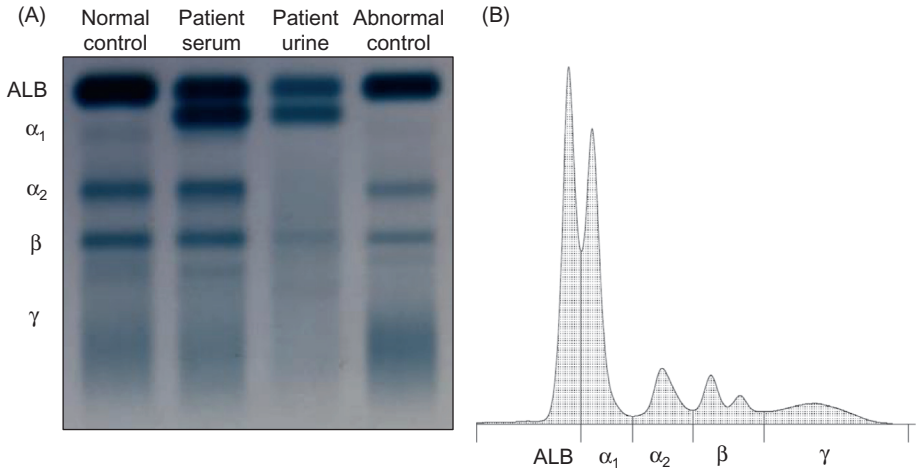


FIGURE 22.2 (A) Serum and urine protein electrophoresis of a patient with bisalbuminemia and normal and abnormal serum controls. (B) The serum protein electrophoresis trace created with a densitometer from the patient with bisalbuminemia. Image credit UVA Clinical Chemistry Fellow Kwabena Sarpong, PhD.

α_2 Region

Haptoglobin (Hp) irreversibly binds free hemoglobin (Hb) as one of its several physiological functions. The Hp protein is synthesized by hepatic parenchymal cells as two heavy chains and two light chains. There are two Hb binding sites per Hp molecule. Only a small amount of in vivo hemolysis is necessary to deplete Hp; for this reason, Hp measurements are clinically useful as a sensitive biomarker of intravascular hemolysis. It may also be used to rule out in vitro hemolysis. After binding to Hb, the larger Hp−Hb complex is less likely to be filtered by the glomerulus, so this binding presumably protects the renal tubules from Hb-induced injury. The Kupfer cells of the liver rapidly clear the complex. Hemolysis does not induce Hp synthesis; however, production is stimulated by inflammation. Although complexation of free Hb appears to be an important function, the Hp−Hb complex is a powerful peroxidase and may cause cytotoxicity to macrophages. In addition, Hp functions as a bacteriostatic agent by binding Hb and deprives iron-requiring bacteria of this element. Hp also inhibits cathepsin B released by phagocytes. Anhaptoglobinemia (OMIM 614081) occurs in approximately 1 in 1000 births and is linked with difficulties in iron (Fe) metabolism and toxicity in patients with chronic hemolysis secondary to malaria.

α_2-Macroglobulin (A2M) is a large, 800-kDa protein synthesized primarily by hepatic parenchymal cells. A2M's major function involves inhibition and complexation of a broad spectrum of proteases. A2M also inactivates kinins and members of the complement and fibrinolytic pathways. A2M is a binding protein for insulin and other small peptides as well as for divalent cations such as zinc. Clearance of A2M is through receptors on hepatic cells. A2M concentrations in children are higher in adults; women tend to have higher concentrations than men. A2M concentrations are increased in the nephrotic syndrome because of induction of this protein's synthesis by the liver, presumably as a compensatory response to maintain oncotic pressure when ALB and other proteins are lost by urinary excretion. A2M synthesis is decreased by the release of IL-1; the protein is a negative APR protein. A2M binds to prostate-specific antigen (PSA) with higher affinity than α_1-antichymotrypsin (ACT). However, the A2M-PSA complex is rapidly cleared from circulation, and 70% of circulating PSA is present in the ACT-PSA complex.

Ceruloplasmin (Cp) is a ferroxidase enzyme synthesized and catabolized by hepatic parenchymal cells. Cp carriers more than 95% of plasma copper. The primary physiologic role for Cp is to facilitate oxidation−reduction reactions in plasma such as the ferric (Fe III) to ferrous (Fe II) conversion of Fe—an essential part of Fe metabolism. The main clinical use for Cp measurements is in screening for Wilson's disease (OMIM 277900), also termed hepatolenticular degeneration. Wilson's disease is an autosomal recessive condition that reduces the activity of an intracellular ATPase. The estimated prevalence of Wilson's disease is 1 case per 30,000 live births in most populations. In the absence of this enzyme, copper is transported to the liver, but the metal is not subsequently bound to Cp. The result is the deposition of copper in hepatic parenchymal cells, in the brain, and in the periphery of the iris also known as "Kayser−Fleischer" rings. Kayser−Fleischer rings appear in approximately 95% of patients with this disorder. Usually, symptoms of Wilson's disease begin in the second or third decades of life, presenting as acute or chronic hepatitis, neurologic symptoms, renal tubular acidosis, or hematologic abnormalities. Cp concentrations are greatly decreased in Wilson's disease. However, Cp is a positive APR protein, so interpretation can be confounded by conditions such as inflammation and pregnancy. The definitive diagnosis of Wilson's disease requires demonstration of increased tissue copper on biopsy and/or analysis of ATP7B gene mutations.

β Region

The β region on SPEP typically contains two distinct bands referred to as β_1 and β_2. The β_1 band is mostly comprised of β-lipoprotein (low-density lipoprotein, LDL) and transferrin (Tf). Very low-density lipoprotein may appear also in the pre-β region. LDL is the most anodal protein in the β_1 region and appears as an anodal "feathery edge." An increase in peak size in this region may indicate the body is not properly breaking down lipids such as seen in conditions of hyperlipoproteinemia or familial hypercholesterolemia. The β_2 band mostly contains complement proteins such as complement component 3 (C3).

β_1 Region

Tf is the most abundant protein in the β_1 band. It is a single-chain protein that binds and transports Fe in circulation. Tf is synthesized in the liver and reticuloendothelial system and has a half-life of approximately 7 days. Tf catabolism is mediated by binding to receptors on hepatic parenchymal cells. Tf is a negative APR protein that is also decreased in chronic liver disease, malnutrition, renal insufficiency, and protein-losing enteropathies. As shown in Table 22.2, Tf is increased with pregnancy, estrogen therapy, and particularly Fe deficiency.

Tf can be measured by immunoassay or estimated using its predictable relationship with total iron-binding capacity (TIBC) according to the following equation:

$$\text{Tf} = 0.70 \times \text{TIBC (mg/dL)} \quad \text{or} \tag{22.i}$$

$$\text{TIBC} = 1.43 \times \text{Tf (mg/dL)} \tag{22.ii}$$

TABLE 22.2 Effects of various hormones and conditions on proteins.

Condition	TTR	ALB	A1AT	AGP	Cp	Hp	Tf	C3	CRP
Inflammation or trauma	D	D	I	I	I	I	D	I(D)(N)	I
Androgens	N	N	I	I	I	I	I	I	N
Estrogens (pregnancy and birth control pills)	N(D)	N(D)	I	N(D)	I	D	I	N	N
Chronic hepatocellular disease	D	D	N(D)	N(D)	N	D	D	N	N or I
Malnutrition	D	D	N(D)	N(D)	N(D)	N(D)	D	N(D)	N
Iron deficiency	N	N	N	N	N	N	I	N	N
Hemolysis, in vivo	N	N	N	N	N	D	N	N	N
Nephrotic syndrome	D	D	D	D	N(D)	I,D	0	N	N
Adrenal glucocorticoids	I	N	N	I	I	I	N	N(I) N	

A1AT, α_1-Antitrypsin; *AGP*, α_1-acid glycoprotein; *ALB*, albumin; *C3*, complement 3; *Cp*, ceruloplasmin; *CRP*, C-reactive protein; *D*, decreased compared with normal; *Hp*, Haptoglobin; *I*, increased compared with normal; *N*, within normal limits; *Tf*, transferrin; *TTR*, transthyretin.

Tf's primary clinical use is for the assessment of Fe status. Normally, approximately 70% of Tf binding sites are saturated with Fe, leaving excess capacity. In Fe deficiency anemia, the Fe concentration decreases, while the Tf protein and TIBC increase. The increase in Tf concentration may lead to an increase in β_1 on SPEP and should prompt investigation into the patient's Fe status if possible.

High saturation of Tf indicates possible hemochromatosis, an autosomal recessive condition, which results in abnormally high absorption of Fe and Fe deposition in the pancreas, heart, liver, joint linings, and gonads. Deposition results in scarring that can lead to progressive tissue injury. Although genetic analysis is the definitive means for diagnosis, increased Fe saturation of Tf may be considered a useful screen. CSF contains two forms of Tf: the form normally found in serum and a desialated isoform termed the "tau protein," Beta Trace Protein or β_2-Tf. The tau protein is abundant in CSF and migrates independently of the intact (sialated) Tf on conventional electrophoresis. Occasionally, there is a clinical question involving the origin of a fluid (e.g., from the nose or ear) and the identification of tau protein by electrophoresis and immunofixation electrophoresis (IFE) with Tf antisera may help determine if the fluid source is CSF. Carbohydrate-deficient Tf is another form of Tf monitored in chronic alcoholics. In nephrology, tau protein is typically termed Beta Trace Protein; high plasma levels of this protein occur in renal failure, inflammation, and tumors, and urine levels have been used to assess residual renal function in dialysis patients.

β_2 Region

C3 is the most abundant protein in the β_2 region of SPEP. C3 has the highest concentration of all complement proteins and represents a functional link between the classical and alternative pathways for the activation of the complement systems. C3 synthesis occurs at several locations, including hepatic parenchymal cells (induced by cytokines and TNF-α), and by fibroblasts and monocytes (induced by endotoxin). C3 is activated by hydrolysis into forms that bind to membranes and other entities, including immunoglobulins. In biliary obstruction, C3 concentrations are proportional to bilirubin increases for reasons that are poorly understood. In focal glomerulosclerosis, increased C3 concentrations are associated with a favorable prognosis. Decreased C3 concentrations increase the risk of infection. Pathologically decreased C3 concentrations may be due to a genetic deficiency or to consumption from autoimmune diseases such as systemic lupus erythematosus (SLE). Electrophoretic variants may be misinterpreted as a monoclonal protein (M-protein) unless definitive identification or quantitation is performed. C3 is a positive APR protein. The appearance of C3 in the β_2 region of quality control material on SPEP may appear degraded, which is attributed to C3 being labile.

γ Region

Immunoglobulin G (IgG), Immunoglobulin A (IgA), Immunoglobulin M (IgM), Immunoglobulin D (IgD), and Immunoglobulin E (IgE) represent the preponderance of protein in the γ electrophoretic region labeled in Fig. 22.1. This region on protein electrophoresis is broad and diffused in appearance. Typically, no demarcated bands should be identified in the healthy population. In general, IgG migrates in the gamma region, IgA migrates to the beta region and the more anodal (early) gamma region, and IgM also tends to migrate toward the anodal

portion of the gamma region. In addition, although some of the immunoglobulins are polymers of similar molecules, all are composed of monomers having two heavy (~55 kDa) chains and two smaller light chains (each ~23 kDa). The G, A, M, D, and E identities of immunoglobulins are defined by the amino acid sequence of the heavy chain, which are signified by the Greek letters γ, α, μ, δ, and ε, respectively. Disulfide bonds covalently bind two like heavy chains; each heavy chain is also covalently bound to a light chain, as indicated in Fig. 22.3. Terminology describing immunoglobulin structure divides the molecule into the Fab fragment (Fig. 22.3), the hinge region joining the heavy and light chains, and the Fc region.

There are two types of light chains for all immunoglobulins. They are known as kappa and lambda and are denoted by the Greek letters κ and λ, respectively. Each κ and λ light chain has a different constant region and a variable region specific for the myriad of antigens that may be encountered. The Fab fragment forms the antigen-binding site, as shown in the upper part of Fig. 22.3. Both light chains for an individual immunoglobulin molecule are either κ or λ; an individual immunoglobulin molecule never mixes one of each. In humans, κ light chains predominate on immunoglobulins in a 3κ:1λ ratio. Free light chains (FLCs) in serum have a different κ:λ ratio of approximately 0.26:1.65. However, recent literature cautions the transference of this ratio on select analyzer platforms without proper verification [5]. The interpretation of the ratio should also be considered in the context of renal status [6]

Immunoglobulins are synthesized mainly in lymph nodes and in blood by plasma cells—a fully differentiated B cell that produces a single type of antibody. A naïve B lymphocyte bears antigen receptors specific for a single chemical structure on its surface. Upon interaction with the target antigen, the cell is activated and proliferates into a clone of identical plasma cells that can secrete clonotypic antibodies with specificity identical to the surface receptor that first triggered the activation leading to the clonal expansion. The interaction of antibodies with cells and complement occurs at the Fc region and varies according to the type and subtype of immunoglobulin heavy chain. The Fab region of the light chain is responsible for interaction with specific antigen, and the hinge region controls the relationship between the Fc and Fab parts of the molecule.

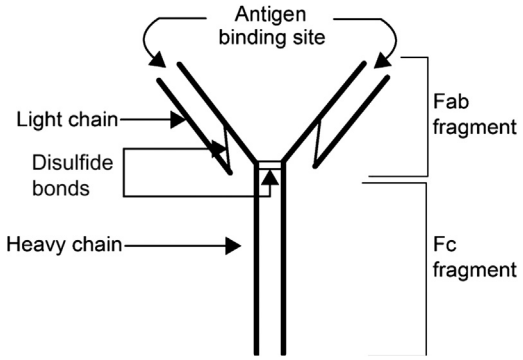


FIGURE 22.3 General structure of immunoglobulin monomers with the Fab and Fc fragments indicated. The hinge region is between the Fab and Fc components.

The γ region of SPEP patterns normally shows homogeneous, diffuse protein staining (Fig. 22.1). However, in the conditions of immunodeficiency, this region can be nearly absent in protein and has a "washed-out" appearance. Examples of such conditions are bone marrow suppression seen in X-lined agammaglobulinemia (OMIM 300755) or "Bruton agammaglobulinemia," an immunodeficiency state characterized by profoundly low concentrations of serum immunoglobulins of all classes. Treatment with plasmapheresis using ALB for fluid replacement can also have a substantial effect on the SPEP pattern, where ALB staining predominates and all of the other regions are rather featureless because of protein removal. Conversely, any reactive or inflammatory process may lead to the increase in immunoglobulins and a polyclonal appearance in the γ region. Patients with autoimmune deficiencies can show abnormal immunoglobulin patterns on SPEP such as polyclonal hypergammaglobulinemia, hypogammaglobulinemia, oligoclonal banding, and occasional monoclonal or triclonal immunoglobulins bands.

IgG is the most abundant immunoglobulin in serum, comprising approximately 75% of total staining in the γ SPEP region. IgG is responsible for long-term physiologic protection from infection accounting for the majority of neutralization of bacterial and viral antigens. Total IgG is comprised of four subtypes termed IgG_1 (65%), IgG_2 (25%), IgG_3 (6%), and IgG_4 (4%). IgG crosses the placenta; neonates have only maternal IgG_1, and adult concentrations of IgG are not reached until approximately 3 years of age. IgG_1 and IgG_3 activate killer monocytes through binding of their Fc receptors on phagocytic cells. The response of many IgG-producing clones to various antigens gives rise to the polyclonal response and homogeneous staining normally observed in the γ region (Fig. 22.1). In particular, increased IgG_4 subclass concentrations give a distinct narrow, smooth, bell-shaped increase in the anodal γ region and can be falsely interpreted as a malignant M-protein [7].

IgA comprises approximately 10%–15% of total immunoglobulins. IgA exists as two subtypes: IgA_1, which predominates (~90%) in the blood, and IgA_2, which predominates (~60%) in mucosal secretions. In the blood, IgA is found mainly as a monomer, while in the mucosal tissues, IgA is almost exclusively a dimer.

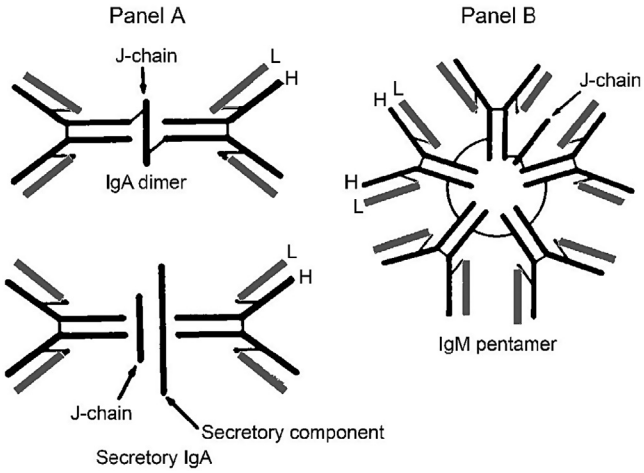


FIGURE 22.4 (A) Structure of immunoglobulin A dimer and the secretory immunoglobulin A subtype containing the secretory component. (B) Structure of the immunoglobulin M pentamer in which the monomers and the J chain are connected by disulfide bonds.

IgA is the major immunoglobulin in tears, saliva, sweat, milk, and also gastrointestinal and bronchial secretions. Secretory IgA has a molecular weight of 380 kDa and consists of two IgA molecules plus a secretory component of 70 kDa and a J chain of 16 kDa (Fig. 22.4 Panel A). The secretory component is thought to enhance solubility in secretions and nonplasma fluids. Increases in IgA frequently observed in liver disease may create a connection or "bridge" of the anodal γ region to the β region on SPEP; this appearance is commonly referred to as β-γ bridging (i.e., beta−gamma bridging).

IgM comprises approximately 10% of total immunoglobulins. IgM is the largest immunoglobulin; as shown in Fig. 22.4 Panel B, the protein exists as a pentamer having a molecular weight of 850 kDa. Along with the five monomeric IgM molecules, the pentamer contains a J chain; disulfide bonds connect the J chain and the monomers. IgM as the transmembrane form is found on the surface of early B lymphocytes and is the most primitive immunoglobulin; after antigen stimulation, some of the B cells differentiate into plasma cells, producing IgM antibodies that are excreted into blood, whereas other cells undergo class switching to form plasma cells of a different class and produce antibodies of that class. IgM is, thus, the first-responder immunoglobulin after antigen assault.

IgD is a monomeric immunoglobulin in plasma and is active on the cell surface of B lymphocytes; its function is incompletely understood. IgD is not measured routinely in clinical laboratories, but antisera are available for particular applications such as IFE (see the section "Immunofixation electrophoresis" for details). IgD is labile on storage, so sample handling and measurement must be completed expeditiously.

IgE is the lowest concentration immunoglobulin and is monomeric in plasma. IgE attaches rapidly to the membranes of mast cells by its Fc region. Cross-linking of two IgE molecules by antigen causes release of histamine, heparin, and vasoactive amines by mast cells. IgE's activity is essential in the allergic reactions associated with urticaria, hay fever, and asthma. Sensitive immunoassays are necessary for quantification, because IgE's plasma concentration is very low ($\leq$214 kU/L for individuals $\geq$18 years old).

C-reactive protein (CRP) has a structure similar to immunoglobulins; key characteristics are shown in Tables 22.1 and 22.2. The term CRP was coined because the protein binds to the polysaccharide coat of *Escherichia coli*. CRP is the most robust positive APR protein, and normal CRP concentrations have a high negative predictive value for infection. The measurement of CRP is recommended over erythrocyte sedimentation rate to detect acute-phase inflammation. The measurement of CRP by highly sensitive methods (hsCRP) can be helpful for estimating the long-term risk of cardiac events. Concentrations of hsCRP in mg/L at $<$1.0, 1.0−3.0, and $>$3.0 carry a relative low, moderate, or high risk of event, respectively. The Justification for the Use of Statins in Primary Prevention: An Intervention Trial Evaluating Rosuvastatin trial demonstrated that patients with elevated hsCRP benefit from statin therapy regardless of LDL concentrations [8].

Standard analytical techniques

Proteins in the clinical laboratory are commonly measured by immunochemistry-based techniques. Specific protein measurement by immunochemistry is based on the in vitro formation of an immune complex by reaction of the protein antigen of interest with a specific antibody (Ab). When the protein antigen is present at high enough concentration (usually in the mg/dL range), immune complexes form an immunoprecipitin. Particle size of the immunoprecipitin is sufficiently large to cause increased turbidity that scatters incident light and results in a decreased signal at the detector. The unbound antigen or Ab does not cause turbidity. Varying the protein antigen concentration when the Ab concentration is held constant forms the immunoprecipitin curve (see Chapter 12, Immunoassays). This curve comprises three distinct areas: (a) Ab excess; (b) area of equivalence; and (c) antigen excess. The immunoprecipitin signal decreases at increased antigen concentrations that passed the area of equivalence; analytical strategies must be in place to detect and avoid antigen excess, because use of this region of the curve will lead to falsely low results at high antigen concentrations.

Turbidimetry and nephelometry

Turbidimetry (immunoturbidimetry) and nephelometry (immunonephelometry) measurements are commonly used to quantify immune-complex precipitates by their ability to interact with incident light. For turbidimetric measurements, the detector is oriented at a straight-on 180-degree angle relative to the incident light source. With this experimental design, the decrease in light intensity is quantified and the amount of light passing through the solution to the detector is inversely proportional to the concentration of protein antigen in the sample. Quantification is accomplished by comparing the signal from the sample with that of a standard curve. Nephelometry detects immune-complex precipitates by their ability to reflect light, a phenomenon termed "Rayleigh light scattering." Because nephelometry measures light scatter, the light detector is oriented at an angle (e.g., 30 or 90 degrees) relative to the incident light source. The amount of light reaching the detector in nephelometry is directly proportional to the quantity of protein antigen in the sample. Quantification is accomplished by comparing the signal from samples with a standard curve.

Radial immunodiffusion

Radial immunodiffusion measurements are performed using gel plates containing a homogeneous distribution of specific Ab for the antigen of interest. Wells are cut in the gel, into which the patient's sample is pipetted. Antigen from the sample diffuses radially until a critical proportion of Ab and protein antigen is achieved (the equivalence point), whereupon a precipitin ring is formed. A standard curve is used, because the area of the precipitin circle (at equilibrium) is directly proportional to the concentration of the protein antigen. An alternative quantification strategy is the Fahey–McElvy modification, in which the diameter is read after a fixed time. The Fahey–McElvy method may produce erroneous results when antigens have multiple forms of differing molecular weight.

Calculated results

The Quotient of ALB (Q albumin) is useful for assessing the integrity of the blood–CSF barrier. The calculation is as follows:

$$Q\ \text{albumin} = 1000 \times \left[\text{CSF albumin}/\text{serumalbumin}\right] \tag{22.iii}$$

All CSF ALB is derived from diffusion or passage into the central nervous system (CNS) from blood, because synthesis occurs only in the liver. Normally, the serum concentration of ALB is approximately 500-fold greater in serum than in CSF. However, disruption of the blood–CSF barrier by inflammation or other process results in increased leakage of ALB in blood into the CSF. The greater the disruption of the blood–CSF barrier, the higher the Q albumin. The Q albumin normal reference interval is <9.

The IgG index is a calculation that "corrects" the relative amount of IgG in CSF for that in serum by normalizing with the Q albumin. The IgG index is calculated as follows:

$$\text{IgG index} = \frac{\text{CSF IgG}/\text{serumIgG}}{Q\ \text{albumin}} \tag{22.iv}$$

Since IgG can be synthesized by plasma cells on either side of the blood–CSF barrier, the IgG index is useful for assessing abnormal IgG production within the CNS. The accepted normal reference interval is 0.30–0.70. Increased values indicate elevated intrathecal synthesis of IgG in the CNS.

Quantitative assessment of proteins

Serum protein electrophoresis

ALB and globulin proteins are two major protein groups in serum. The SPEP test separates these two protein groups on the basis of size and charge properties into the six major regions: ALB, α_1, α_2, β_1, β_2, and γ. The gel matrix most commonly utilized in this technique is agarose; however, cellulose acetate may also be used, but tends to show lower resolution. SPEP performed on an agarose system is typically buffered at pH 8.6. After electrophoresis, the proteins are fixed in the gel that is bonded on a clear Mylar plastic backing. The SPEP pattern is visualized using a general dye-binding protein stain such as Amido black or Coomassie blue. The gel is further dehydrated and scanned.

Most gels may hold up to 30 patient samples. Each gel should include two levels of quality control to be assessed in the same manner as a patient specimen. Densitometry is performed with instrumentation that scans the SPEP pattern, as shown in Fig. 22.5. Incident light passes through the clear Mylar backing of the fixed and stained gel to the detector. However, areas of the gel containing fixed and stained proteins block the incident light. The relative intensity of the incident light at the detector is plotted versus gel position to form the pattern. Software in the densitometer quantifies the proportion of total protein staining represented by each SPEP region. It is important to note that following electrophoresis, fixation, and staining, a direct visual inspection of the gel is an essential component of protein electrophoresis studies, because subtleties occur, which are not resolved on standard densitometric scanning. M-proteins or other

abnormal findings will lead to additional and more sensitive testing such as subtyping by IFE or immunosubtraction. Through a case example, these techniques will be expanded upon in the subsequent text of this chapter.

UPEP is based on the same separation principles as SPEP. However, the concentration of protein in urine specimens is normally far lower than in serum; therefore preconcentration of samples is required before UPEP. Random or 24-hour urine specimens may be used in the analysis. Densitometric UPEP patterns are sometimes helpful for quantifying protein excretion and for monitoring the presence of monoclonal banding, particularly the excretion of κ or λ light chains in urine (e.g., Bence Jones protein defined as a monoclonal or immunoglobulin light chain found in the urine, with a molecular weight of 22–24 kDa) [9]. CSF electrophoresis is usually performed like SPEP; however, more sensitive IEF (i.e., isoelectric focusing) methods are available. CSF specimens generally have low protein content, so samples require either preconcentration before analysis or use of more sensitive stains (e.g., silver stain).

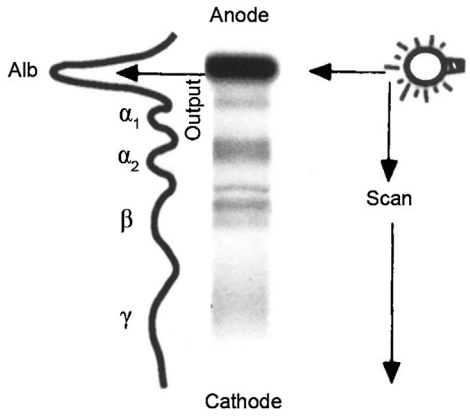


FIGURE 22.5 Sketch of the experimental design for densitometry. A scan of a normal SPEP is indicated. As the incident source scans the gel, the intensity of light passing through the gel and reaching the detector varies according to the staining intensity, resulting in the output shown.

Case example

A 70-year-old African American male presents to the emergency department with complaints of fatigue, spinal column pain, and frequent emesis. The care team orders complete cell counts (CBC), comprehensive metabolic panel (CMP), urinalysis, and protein electrophoresis. His select CBC, CMP, and urinalysis results are listed in Table 22.3, and his SPEP is shown in Fig. 22.6.

This is a classic example of a patient presenting with nonspecific findings such as fatigue and bone pain. Laboratory results of decreased Hb and increased blood urea nitrogen, creatinine, total protein, and calcium are all concerning for someone with a plasma cell disorder. Additional diagnostic testing was performed, and a diagnosis of multiple myeloma (MM) was determined.

Immunofixation electrophoresis

IFE is able to characterize further or subtype M-proteins identified on SPEP such as the one identified in the recent case example. This technique takes advantage of Ab specificity for antigens and their ability to fix (e.g., immunoprecipitate)

TABLE 22.3 Results of clinical laboratory testing for case example.

Select laboratory tests		Result	Reference interval	Units
CBC	WBC	5.6	4.8–10.8	K/μL
	RBC	2.6	4.0–5.4	M/μL
	Platelets	162	145–499	K/μL
	Hemoglobin	8.6	12.0–16.0	g/dL
	Hematocrit	25.3	36.0–47.0	%
CMP	BUN	40	7–18	mg/dL
	Creatinine	4.5	0.5–1.2	mg/dL
	Calcium	10.2	8.5–10.2	mg/dL
	TP	10.5	6.1–8.0	g/dL
	ALB	2.7	3.5–4.8	g/dL
Urinalysis	Protein	28.0	< 30.0	mg/dL

ALB, Albumin; *BUN*, blood urea nitrogen; *CBC*, complete cell counts; *CMP*, comprehensive metabolic panel; *RBC*, red blood cells; *TP*, total protein; *WBC*, white blood cell.

proteins in a gel matrix. IFE is performed after the specimen is subjected to electrophoresis in the same process as SPEP but in multiple lanes. After electrophoresis is complete, each isolated electrophoretic lane is then overlaid with specific antiserum that reacts with the protein antigen(s) of interest. The antiserum precipitates the specific protein of interest, fixing it in the gel matrix. After washing away unreacted reagent and unfixed proteins, the gel is stained and the fixed immune precipitates are identified. The additional protein from the Ab-antigen IFE reaction enhances sensitivity for detection by magnifying the intensity of staining several fold.

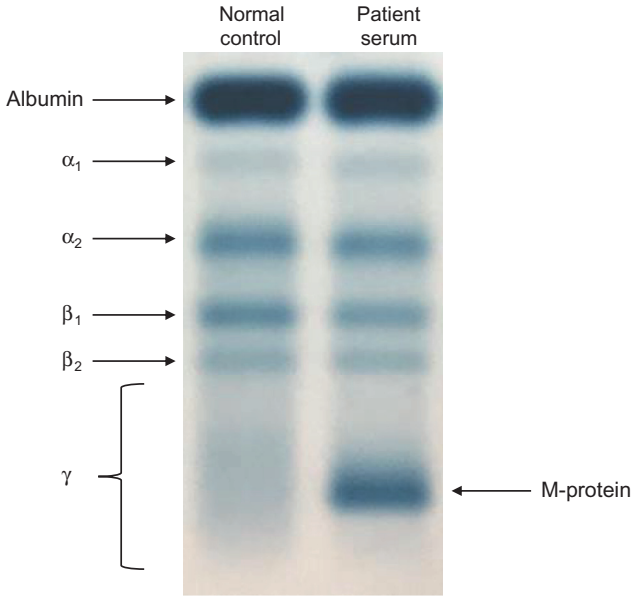


FIGURE 22.6 Control (normal) serum protein electrophoresis pattern *(left lane)* and serum protein electrophoresis of the patient's serum from the case example showing a monoclonal gammopathy *(right lane)*.

IFE of serum is typically performed using specialized gels in which electrophoresis is performed simultaneously in six adjacent lanes for each specimen. Following separation and placement of the template cutout, the first lane is overlaid with a general protein fixative and may be used as a reference. Each of the second through sixth lanes is overlaid with respective antisera against the IgG, IgA, and IgM heavy chains, and then κ and λ light chains, respectively. The IFE gel from the case example shows the characterization of the M-protein as an IgG λ (see Fig. 22.7).

IFE of urine requires electrophoresis of the patient's preconcentrated urine specimen in multiple lanes. Generally, the same six-lane specialized gels (Fig. 22.8) as for serum IFE are used. Since the focus of urine IFE is often detection of Bence Jones proteinuria, a general fixative is typically overlaid in the first (leftmost) lane; free κ (unbound to heavy chain), total κ, free λ, and total λ are overlaid in the second through fifth lanes, respectively. The last lane can be used to assess urine excretion of any heavy-chain components identified on serum IFE.

IFE of CSF typically requires specimen concentration or sensitive stains, and uses the same specialized gels as are used for serum IFE. CSF IFE is usually performed to assess the presence of oligoclonal banding, defined as multiple bands of IgG identity migrating in the γ region. The strategy for CSF IFE typically includes a general fixative in the first lane and IFE with IgG, κ, and λ antisera in separate lanes. Fig. 22.9 shows that IFE of CSF enhances the ability to detect oligoclonal bands due to its greater sensitivity compared with conventional electrophoresis and staining (see the section "Multiple sclerosis and cerebrospinal fluid electrophoresis").

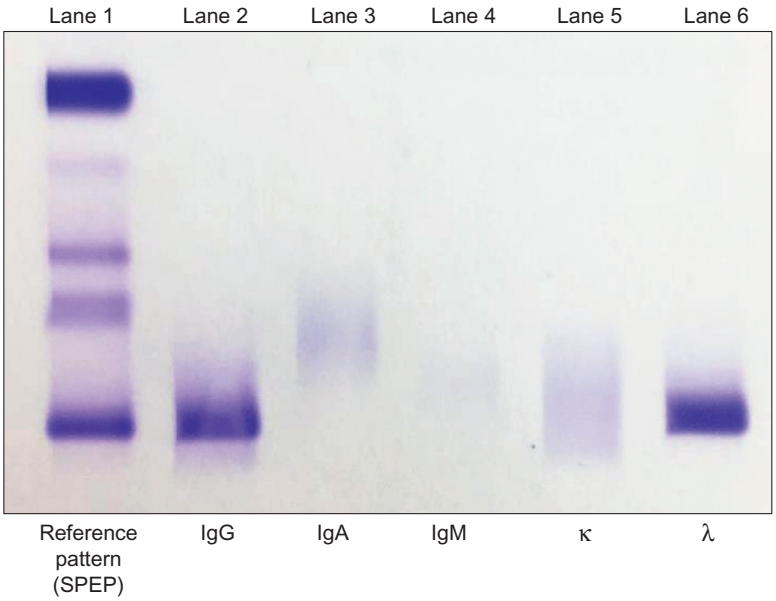


FIGURE 22.7 Six-lane immunofixation (immunofixation electrophoresis) gel of the case example in which a general fixative is applied to the *leftmost lane* (reference pattern). Each of the remaining lanes is overlaid with specific antisera for γ (immunoglobulin G), α (immunoglobulin A), and μ (immunoglobulin M) heavy chains, and κ and λ light chains as indicated. In lanes 2–6, only the protein of interest is fixed for staining and visualization. Note that the identity of the monoclonal protein is immunoglobulin G, with λ light chain.

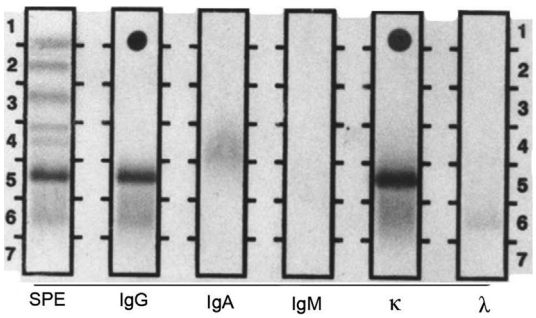


FIGURE 22.8 Another typical six-lane immunofixation (immunofixation electrophoresis) gel, in which a general fixative is applied to the *leftmost lane*. Each of the remaining lanes is overlaid with specific antisera for γ (immunoglobulin G), α (immunoglobulin A), and μ (immunoglobulin M) heavy chains, and κ and λ light chains as indicated. In lanes 2–6, only the protein of interest is fixed for staining and visualization. Note that the identity of the monoclonal protein is immunoglobulin G, with κ light chain.

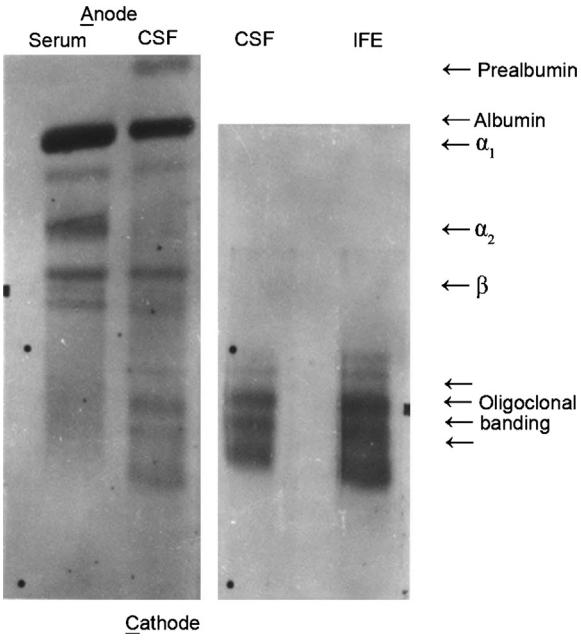


FIGURE 22.9 The *leftmost lanes* show serum and cerebrospinal fluid electrophoresis patterns for a patient with multiple sclerosis; note the presence of oligoclonal banding in the cerebrospinal fluid. The prealbumin band stains prominently in the cerebrospinal fluid lane. In the *rightmost two lanes*, the cerebrospinal fluid was immunofixed with immunoglobulin G antisera. Note the greater intensity of the immunofixed lanes. Cerebrospinal fluid applied to the *rightmost lane* was subject to more preconcentration prior to electrophoresis.

Capillary zone electrophoresis

Capillary zone electrophoresis (CZE) is an efficient, sensitive, and versatile analytical method that is based on combining zone electrophoresis, isoelectric-focusing electrophoresis, and isotachophoresis (see Chapter 8, Chromatography and electrophoresis). CZE systems have

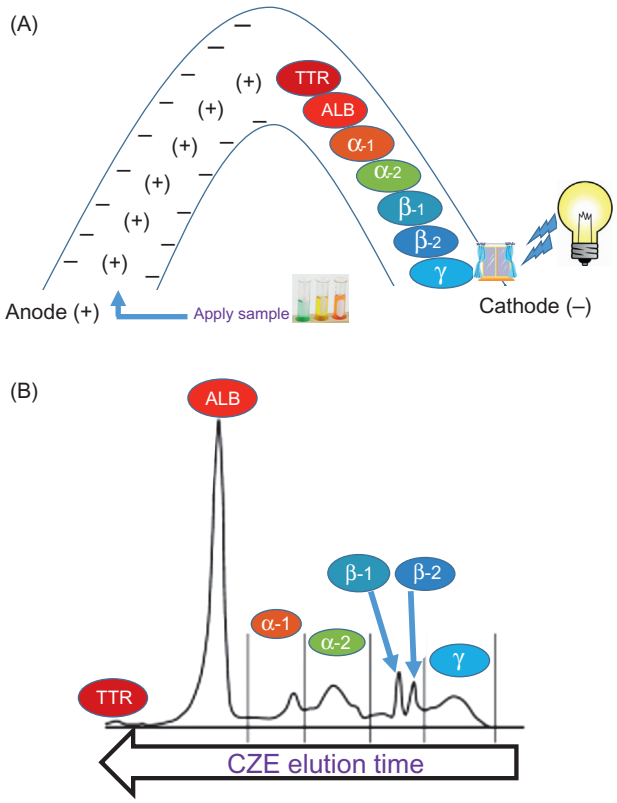


FIGURE 22.10 *Panel A:* Sample is applied into the capillary tube and migrates through the charged environment, with the positively charged proteins passing through the column most easily. Detection of the proteins is possible through a window in the column. *Panel B:* The protein pattern resulting from capillary zone electrophoresis showing the direction of elution from the column. Note that it is similar to the serum protein electrophoresis pattern displayed in Fig. 22.5.

been investigated for clinical use for several decades, and there are various instruments available for assessing protein regions in serum and urine. Rather than a gel-based technology, CZE utilizes a liquid-based system that separates proteins in a strong negatively charged environment, as shown in Fig. 22.10A. As shown, the sample is aspirated into the anodal end of a fused silica capillary, where the negative charge on the interior of the capillary applies a strong endosmotic flow of cations toward the cathode. The proteins travel through the capillary at a rate proportional to the intensity of their net positive charge with a flow of negatively charged proteins such as the ALB moving slowest and positively charged proteins such as immunoglobulins moving through the column more quickly (Fig. 22.10A). As the protein fractions exit the capillary system, they are detected by absorbance, commonly at 200–215 nm, through a window in the column. This results in the electropherogram shown in Fig. 22.10B. There is a strategy to identify M-proteins in which specific antisera are added to the specimen of interest. The specific antisera remove the M-protein targeted by

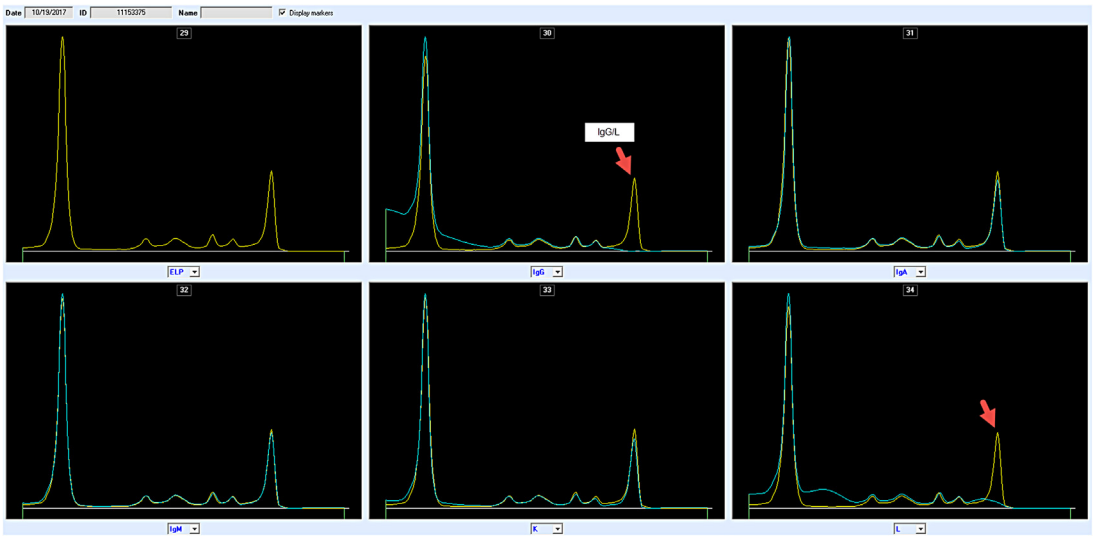


FIGURE 22.11 Typical capillary zone electrophoresis immunosubtraction panels (Sebia, Capillarys) from the case example in which a general capillary pattern is depicted in the *leftmost lane* (reference pattern). Each of the remaining lanes is treated with antibodies against γ (immunoglobulin G), α (immunoglobulin A), and μ (immunoglobulin M) heavy chains, and κ and λ light chains. Then capillary zone electrophoresis is performed to determine which heavy and/or light chains are removed/subtracted. The reference pattern is applied to panels 2–6. Electrophoretic shifts of bounded proteins can be seen in the albumin region of panels 2 and 6. This characterizes the monoclonal protein as immunoglobulin G, with λ light chain.

the antisera by precipitation, which is eliminated from the sample. This technique is commonly referred to as immunosubtraction. CZE before and after immunosubtraction eliminates the monoclonal band, and measurement of the area difference allows quantification of the precipitated M-protein. The CZE trace and immunosubtraction for the case example are shown in Fig. 22.11.

Automated CZE methods are gradually being adopted in the clinical laboratory. Recently, an international survey of 774 laboratories from 38 countries revealed 31.4% ($n = 238$) are currently utilizing CZE for serum protein measurement versus the remainder, who predominantly perform analysis by agarose gel electrophoresis [10]. The migration to CZE from SPEP is likely due to CZE offering relatively straightforward automation, generally improved precision, and overall faster turnaround times than conventional SPEP. However, CZE measurements do not provide a physical gel for visual inspection, although computer generated or "digital gels" can be provided by current software. In addition, the absorbance detection of CZE results in various interferences and the immunosubtraction technique requires additional training and experience. The overall cost per test is more expensive than traditional gel electrophoresis. These limitations may delay adoption of CZE, and labs will need to weigh their options such as volumes and reimbursements.

Mass spectrometry

The use of mass spectrometry to identify, isotype and quantitate M-proteins may be on the horizon for the clinical laboratory. Mass spectrometry has shown better analytical sensitivity and specificity in the detection of M-proteins than SPEP or CZE. Recent studies by Mills et al. [11] have demonstrated the utility of M-protein assessment using nanobody enrichment coupled to matrix-assisted laser desorption ionization-time of flight also known as "MASS-FIX." As of 2018, MASS-FIX is offered clinically as a laboratory-developed test at a single referral lab in the United States. The method is capable of performing rapid analysis (<1 minute per patient), but the 2.5-hour sample preparation is a recognized limitation.

Mass spectrometry may also assist in monitoring therapeutic monoclonal antibody (t-mAb) treatments in patients with MM. Patients with comigrating IgG κ malignant M-proteins and IgG κ therapeutic monoclonal antibody (t-mAb) on IFE could benefit from the specificity offered by mass spectrometry. A recent pilot study demonstrated mass spectrometry was able to differentiate accurately 100% of samples ($n = 192$) containing t-mAb, an endogenous M-protein, or both regardless of migration patterns observed on IFE. This work highlights the growing use of mass spectrometry in this area and the potential elimination of false-positive results due to t-mAb interferences [12]. In the future, mass spectrometry may be the method of choice for protein measurement. However, upfront instrument costs as well as the required instrument expertise may keep SPEP and CZE as frontline tests for several years to come.

Serum free κ and λ light chains in serum and urine

Free κ and λ light chains may appear on protein electrophoresis at high concentrations. There are four major indications for FLC testing in the evaluation and management

of MM and related clonal plasma cell disorders: screening, prognosis, monitoring for recurrence, and monitoring therapeutic response. Screening for clonal plasma cell disorders (including MM) by simultaneous analysis of SPEP and examination of the serum FLC assay negates the need for screening by 24-hour urine studies. However, it is important to note that once diagnosis of a plasma cell disorder is made, 24-hour urine studies are required for all patients. The serum FLC assay results are of major prognostic value in virtually every plasma cell disorder, including monoclonal gammopathy of undetermined significance (MGUS), smoldering myeloma, active myeloma, immunoglobulin light-chain amyloidosis, and solitary plasmacytoma. Quantitative monitoring of patients with oligosecretory plasma cell disorders (including patients with AL amyloidosis, oligosecretory myeloma, and nearly two-thirds of patients who had previously been deemed to have nonsecretory myeloma) is accomplished by examination of results from serum FLC measurements. In addition, measurement of FLC is essential in acute leukemia patients and patients with oligosecretory myeloma. Finally, serum FLC measurement is a requirement for monitoring and evaluating clinical response, according to the *International Myeloma Working Group (IMWG) Uniform Response Criteria for Multiple Myeloma* [13].

β_2 Microglobulin

As a tumor marker, β_2-microglobulin is used to provide independent information about prognosis and tumor burden. β_2-microglobulin is not a diagnostic biomarker, because it is nonspecific for malignancy or other disease processes.

Three staging criteria for prognostic interpretation of β_2-microglobulin values are available from "*International Staging System for Multiple Myeloma*" [14].

1. Stage I criteria require serum β_2-microglobulin to be $<$3.5 mg/L (296 nmol/L) and serum albumin $>$3.5 g/dL (35 g/L or 532 μmol/L; median survival 62 months).
2. Stage II criteria fits neither stage I or III (median survival 45 months).
3. Stage III criteria requires serum β2-microglobulin to be $>$5.5 mg/L (465 nmol/L; median survival 29 months).

Other indications to measure β_2-microglobulin are to assess tubular dysfunction in renal tubular disease.

Interferences

Analytical interferences are common to all areas of the clinical laboratory. Interferences may be due to a natural or pathological process or may be from some exogenous compound such as medications or medical therapies. Multiple interferences for both agarose gels and capillary-based methods for protein determination have been well characterized by McCudden et al. [15]. However, unrecognized interferences are a common source of laboratory error and may lead to additional testing, inappropriate treatments, or even misdiagnosis. Therefore it is imperative for the laboratorian to be aware of the type of interferences currently identified with their technology and how to mitigate or eliminate them if possible. Table 22.4 summaries several interferences that can be encountered during review of protein electrophoresis.

Endogenous interferences

Many common interferences that affect SPEP and CZE are endogenous. The two most frequent patterns identified are fibrinogen and hemolysis. It is important to recognize the various sources and develop strategies to avoid false interpretations. A commonly encountered analytical interference on SPEP and CZE is fibrinogen. Fibrinogen is a large, 340-kDa glycoprotein readily converted into fibrin by thrombin in the final steps of the coagulation cascade. Fibrinogen is typically removed from serum specimens, although patients with coagulation disorders or patients on anticoagulation therapy may contain traces of fibrinogen. On occasion, plasma samples may also be submitted for SPEP or CZE analysis. Specimens containing fibrinogen will have an abnormal band that migrates to the β/γ region. Such a band virtually always requires an IFE or immunosubtraction, to avoid misidentifying the band as an M-protein. If the analyzed specimen was heparinized plasma, diligent use of either of these techniques (IFE or immunosubtraction) can help confirm possible abnormalities seen on SPEP or CZE.

Hemolysis is a widely characterized interferant in the clinical laboratory. Lysed or "hemolyzed" red blood cells release high cellular concentrations of Hb, potassium, magnesium, iron, phosphate, lactate dehydrogenase, and aspartate aminotransferase. Hb causes a spectral interference in most absorbance-based methods and may be identified by visual inspection of the plasma or serum sample, or by the generated hemolysis indices reported by most automated chemistry analyzers. On SPEP, the α_2/β_1 regions may show discrete bands attributed to sample hemolysis. Similar to fibrinogen interference, specimen examination or an IFE should be performed to avoid false interpretation of the abnormal pattern. Causes of hemolysis may be due to various in vivo or in vitro processes.

Elevated serum concentrations of polyclonal IgG_4 subclass may mimic or resemble an M-protein. This interference may also be referred to as monoclonal gammopathy mimicry. The endogenous increase is due to IgG_4-releated disease (IgG_4-RD), a syndrome of unknown etiology. On SPEP and CZE, a characteristic focal band can be seen bridging the β/γ regions of patients with IgG_4-RD. The pattern can appear polyclonal on IFE or immunosubtraction

TABLE 22.4 List of common interferences encountered during protein electrophoresis.

Interference	Method affected	Action for resolution	Frequency
Fibrinogen	SPE/IFE (both capillary and agarose gel methods)	Thrombin treatment, ethanol precipitation, preabsorption of antisera	Common
Contrast/radio-opaque dyes	Capillary electrophoresis	IFE is unaffected	Uncommon
Antifungals (5-fluorocytosine)	Capillary electrophoresis	IFE is unaffected	Uncommon
Antiboiotics	SPE (capillary and agarose gel)	IFE is unaffected	Uncommon
Hemolysis	SPE (capillary and agarose gel)	IFE is unaffected	Common
Heterophilic antibodies	IFE (capillary and agarose gel)	Clinical awareness/education	Rare
Polyclonal IgG_4 subclass mimics	IFE (capillary and agarose gel)	Clinical awareness/education	Rare
Monoclonal therapies	SPE/UPE/IFE (capillary and agarose gel)	Clinical awareness/education, migration shift assays, and mass spectrometry	Rare, but may become significant in the future

IFE, Immunofixation electrophoresis; *IgG*, immunoglobulin G; *SPE*, serum protein electrophoresis; *UPE*, urine protein electrophoresis.
Modified from reference C.R. McCudden, J.F.M. Jacobs, D. Keren, H. Caillon, T. Dejoie and K. Andersen, Recognition and management of common, rare, and novel serum protein electrophoresis and immunofixation interferences, *Clin. Biochem.* **51**, 2018, 72–79,

with both κ and λ bands staining similarly. This may lead to confusion with clonality, and IEF should be performed to differentiate. Human antianimal antibodies (HAAAs) and heterophile antibodies are known but relatively rare endogenous interferences on SPEP [15]. Human antimouse antibodies are the most common subclass of HAAAs. Exposure to farm animals or medical treatments with animal proteins may elicit an immune response and formation of HAAAs. No previous exposure is necessary for the development of a heterophile antibody. Heterophile antibodies are IgM class immunoglobulins. Although HAAAs are nonpathological, they may give the false appearance of an M-protein and may lead to incorrect classification as MGUS.

Endogenous proteins in the β and γ region can also be possible sources of interference and confusion. Proteins in the β_1 and β_2 region may mask or hide M-proteins. Patients presenting for the first time with clinical indications of MM but no demarcated band on SPEP or CZE may require an IFE or immunosubtraction to rule out a β region, M-protein. Unfortunately, M-proteins in this region can also be difficult to quantitate, and these limitations should be shared via comments or alerts to clinicians. As described early in this chapter, proteins such as Tf can also mimic an M-protein in the β_1 region and may lead to additional testing such as IFE. Finally, CRP is commonly misinterpreted for an M-protein. During an inflammatory response, CRP can show up as a band in the γ region; however, since CRP does not react with heavy-chain or light-chain antisera, the band is absent on IFE.

Exogenous interferences

It has been observed that exogenous compounds, such as radiocontrast dyes, medications, and more recently m-tAb, can affect the interpretation of SPEP or CZE. These compounds may interfere with the spectral absorbance in CZE, leading to a falsely abnormal band, or directly affect the interpretation by the presence of a quantifiable m-tAb. The list of exogenous interferences continues to grow with the gamut of current and new drugs, supplements, and treatments being given to patients.

Radiopaque chemicals or "agents" absorb at $\sim$200 nm, the same wavelength utilized by CZE. This often leads to the detection of an additional spike or band, which may be concerning for the presence of an M-protein. The α_2 or less commonly the β_2 regions are mostly affected by the radiopaque agents. Fortunately, the abnormal pattern can be further investigated by IFE or immunosubtraction, but the additional analysis is expensive and can take away important resources (e.g., time). Laboratories performing CZE may want to educate clinicians about the possible interference and make further recommendations about timing appropriate specimen collection whenever possible. These radiopaque chemicals do not interfere with conventional SPEP agarose gels.

Some antifungal and antibiotic medications can elicit abnormal patterns on CZE. Interference from 5-fluorocytosine on CZE may be seen at the end of the γ region of patients with immunosuppression and renal failure.

Similar to the radiopaque chemicals, 5-fluorocytosine absorbs around the 200-nm wavelength utilized in CZE. This pattern increases the suspicion of an M-protein and IFE or immunosubtraction should be performed to confirm. An abnormal band between α_2 and β_1 is associated with piperacillin. Other medications such as ceftriaxone and sulfamethoxazole produce abnormal bands or spikes but appear on CZE in regions less associated with M-proteins (e.g., prealbumin and ALB).

Interference from novel monoclonal therapies is a growing concern for clinicians and laboratorians. Treatments with targeted monoclonal antibodies are becoming common for patients with MM and other plasma cell disorders. Therapeutic doses of daratumumab, siltuximab, elotuzumab, and newer t-mAbs can appear on SPEP and CZE as a small M-protein. Most t-mAbs identified by IFE or immunosubtraction are IgG_1 κ. The half-life of many of these biologic drugs can cause interferences several weeks after infusion and may be misleading to the clinician who might be expecting partial or complete response from therapy.

Several approaches may help to alleviate t-mAb interference. The most practical method involves collaboration with other clinical departments. For example, the hospital pharmacy may provide lists of patients receiving immunotherapy to the laboratory or the laboratorian performing signouts can examine the medical record when bands suspicious for immunotherapy are present on SPEP or CZE. This allows the laboratory to track and provide comments where appropriate. An additional approach involves the use of a commercially available methodology that utilizes antidaratumumab antibodies to decipher comigrating malignant M-protein from daratumumab. Similar approaches are being investigated with other t-mAbs; however, the rate of t-mAb development and wide clinical use is expected to increase and it will be challenging for vendors and laboratorians to keep up. Another effective approach to this complex solution involves the use of mass spectrometry analysis. The exact mass of the specific t-mAb can alleviate concerns of M-protein identify and more importantly help to eliminate other endogenous and exogenous interferences. However, these methods are not considered mainstream yet and until a better approach is readily available to minimize or remove m-Ab interference, and cooperation between the clinical laboratory and medical community will be essential to patient care.

Protein analysis and interpretation

Protein electrophoresis examples

Fig. 22.12 displays selected protein patterns that may be encountered in clinical practice. The normal pattern is shown in the first lane of Fig. 22.12. As with any clinical assay, inclusion of appropriate controls is a fundamental part of SPEP and UPEP interpretation.

The glomerular UPEP pattern in Lane A is typically quite featureless except for the obvious ALB band and a faint band in the β_1 region. Mechanistically, this pattern results when negatively charged ionic repulsion at the basement membrane of the glomerulus is compromised and allows ALB and other low-molecular-weight, negatively charged proteins to pass through the glomerulus, overwhelm the protein resorbing mechanisms, and pass into the urine. This pattern typically depicts ALB and TRN in the β_1 region, but λ and κ light chains can migrate in the β region too. If there is high clinical suspicion and increased serum FLC, a UPEP IFE may be performed. Nonselective proteinuria (Lane B) typically indicates a lack of the kidney's ability to filter and

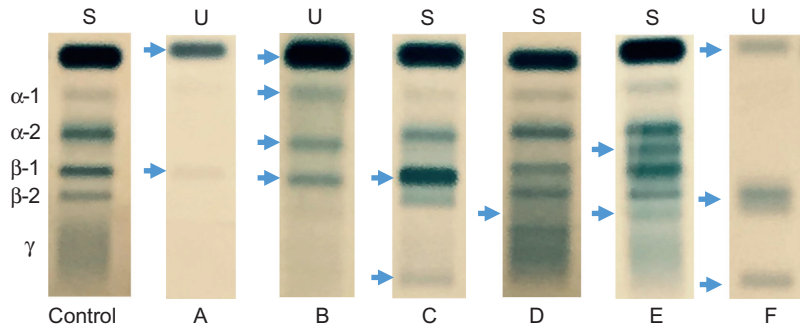


FIGURE 22.12 Examples of various protein electrophoresis patterns encountered in clinical practice. U represents urine and S represents serum. *Control Lane:* serum protein electrophoresis of Normal Serum. *Lane A*: glomerular urine protein electrophoresis pattern. Arrows point to characteristic albumin band and faint band in β_1 region (typically transferrin). *Lane B*: nonselective urine protein electrophoresis pattern, showing excretion of nearly all serum protein fractions (highlighted by the arrows). *Lane C*: serum protein electrophoresis, showing biclonal band. The arrow in the β_1 region points out a monoclonal-protein comigrating with proteins of this region. The arrow in the γ region highlights a second, faint band. *Lane D*: serum protein electrophoresis, showing $\beta-\gamma$ bridging (*arrow*). Note there is a faint band in the γ region just to the γ side of the arrow. *Lane E*: serum protein electrophoresis, showing a hemoglobin band migrating between the α_2 and β_1 regions (*arrow*). There is also a fibrinogen band (*arrow*) due to incomplete coagulation of the specimen. *Lane F*: a complicated urine protein electrophoresis showing albumin excretion due to glomerular injury, a band in the γ region that is an intact immunoglobulin from serum (extremely high, >4 g/dL) that "overflowed" into urine, and bands in the β region due to Bence Jones proteinuria.

TABLE 22.5 Association of the various immunoglobulin subtypes with multiple myeloma.

Protein	Frequency of association with MM, %	Bence Jones proteinuria, %	Mean age of occurrence (years)
IgG	50	60	65
IgA	25	70	65
IgD	2	100	57
Free light chains only	20	20	56
None detected	<1	0	-
Bi-clonal	1	-	-
IgM	1	50–100	-

IgA, Immunoglobulin A; *IgD*, immunoglobulin D; *IgG*, immunoglobulin G; *IgM*, immunoglobulin M; *MM*, multiple myeloma.

resbsorb both small- and large-molecular-weight proteins. In addition, the negative charge repulsion mechanism to conserve ALB is compromised, and virtually all fractions of the SPEP pass into urine. Lane C shows a biclonal monoclonal gammopathy pattern (see arrows). As shown in Table 22.5, approximately 1% of MM patients demonstrate a biclonal banding pattern. These bands should be characterized by IFE for monitoring. β-γ bridging (arrow Lane D of Fig. 22.12) is frequently observed in liver disease patients. Note there is also a faint band in the γ region that is an indication for performing IFE. Lane E is an SPEP pattern, showing two possible misleading interferences highlighted by the arrows. First, there is an Hb band migrating between the α_2 and β_1 regions (upper arrow). There is also a fibrinogen band (arrow right below the β_2 region) due to incomplete coagulation or collection of the incorrect specimen type. IFE of this SPEP would identify no corresponding M-proteins. Lane F is a complicated UPEP, showing ALB excretion due to glomerular injury; in addition, there is a band in the γ region that is an intact immunoglobulin from serum (extremely high, >4 g/dL) that "overflowed" into urine, and bands in the β region due to the Bence Jones proteinuria. The legend of Fig. 22.12 details the characteristic findings for each.

Acute-phase response

APR proteins respond to cytokine release and physiologic stresses such as inflammation, infection, trauma, myocardial infarction, or tumors. The increase or decrease in protein concentration in these situations leads to their further characterization. Proteins that increase with such stresses are termed "positive APR proteins"; those that decrease are termed "negative APR proteins." The APR classification of several proteins is listed in Table 22.1. "Acute" is actually a misnomer, because the APR proteins also persistently respond in chronic conditions such as rheumatoid arthritis and SLE. Many other conditions and therapies may affect the APR proteins and the corresponding SPEP pattern, including endocrine status and administration of steroid hormones, as shown in Table 22.2.

Monoclonal gammopathies

M-proteins are detected in approximately 5% of patients older than 50 years and in approximately 8% of patients >70 years. Once an M-protein is detected, characterization and temporal monitoring of the specific protein are important for clinical care. Table 22.6 lists laboratory and other investigative tests useful for screening and diagnosis of MM. Laboratory tests include SPEP and UPEP, IFE or immunosubtraction in serum and urine, serum FLC measurement, as well as routine laboratory studies including CBC, serum creatinine, electrolytes, calcium, lactate dehydrogenase, and β_2-microglobulin [16]. SPEP and IFE are considered the first-line tests for the workup of MM. If MM is suspected, a bone marrow sample should be collected and examined. If an M-protein is detected through SPEP, UPEP, or pathological FLC ratio and the plasma cell count is higher than or equal to 10%, a diagnosis of MM is made [15]. In addition, if any of the following three criteria are met, a diagnosis of MM is made [16]:

- 60% or greater clonal plasma cells on bone marrow examination;
- serum FLC ratio of 100 or greater of either involved or uninvolved κ or λ light chain, provided the absolute concentration of the involved light chain is at least 100 mg/L (the "involved" FLC, either κ or λ, is the one that is above the normal reference range; the "uninvolved" FLC is the one that is typically in, or below, the normal range); and

TABLE 22.6 Tests and procedures for diagnostic workup for multiple myeloma.

	Test
Blood	Serum protein electrophoresis (agarose gel or capillary) and immunofixation or immunosubtraction Serum immunoglobulins quantitative Serum free light-chain assay Total serum protein, serum albumin, creatinine, calcium, electrolytes, lactate dehydrogenase, and β2-microglobulin Hemoglobin, white blood cell count, differential count, and platelet count
Urine	Urine protein electrophoresis (agarose gel or capillary) and immunofixation 24-h urine for total protein and light chains
Bone marrow	Aspirate and biopsy for plasma cell count, morphology, and amyloid[a] Cytogenetic evaluation and fluorescence in situ hybridization for the detection of del 13, del 17p13, t(4;14), t(11;14), t (14;16), 1q +
Bones	Skeletal survey (conventional X-ray) or low-dose CT scan without contrast
Whole body	MRI and PET-CT Tissue biopsy for solitary or extraosseous plasmacytoma

CT, Computed tomography; *MRI*, magnetic resonance imaging; *PET-CT*, positron emission tomography-computed tomography.

[a]*Useful under some circumstances.*
Modified from S.V. Raijkumar, M.A. Dimopoulos, A. Palumbo, et al., International myeloma working group updated criteria for the diagnosis of multiple myeloma, *Lancet. Oncol.***15** (2014) 538–548.

- more than one focal lytic lesion on magnetic resonance imaging (MRI) that is at least 5 mm or greater in size.

MM is also diagnosed in patients with $<10\%$ plasma cells that also have an M-protein of more than or equal to 3 g/100 mL. In patients with nonsecretory MM, the diagnosis is dependent on the bone marrow findings of $>30\%$ bone marrow plasma cells or plasmacytoma detection [16,17].

When plasma cell clones multiply beyond normal control, an M-protein can be identified on SPEP as a darkly staining area, frequently (but not exclusively) in the γ region, as illustrated in Fig. 22.13. M-protein disorders can be conceptualized as a disease continuum spanning from the asymptomatic "MGUS" condition, to "smoldering" myeloma, which has some features of malignancy, all the way to the full-blown MM malignancy. Criteria and definitions of this continuum are listed in Table 22.7 [18]. Pathological findings associated with M-proteins include amyloid light-chain amyloidosis (termed "primary amyloidosis" or AL amyloidosis) and light-chain deposition disease (LCDD). Monoclonal plasma cell disorders include IgG, IgA, IgM, and occasionally IgD or IgE classes; also the disorder can involve AL amyloidosis, Waldenström macroglobulinemia (IgM), MGUS, and B-cell lymphoma. In addition, monoclonal gammopathy may also secrete κ or λ light chains only, without an associated heavy chain component. In these cases, the measurement of serum FLCs is particularly important for assessment of patients; Table 22.8 lists criteria for the continuum of light-chain M-protein that includes MGUS, idiopathic Bence Jones protein, and light-chain MM.

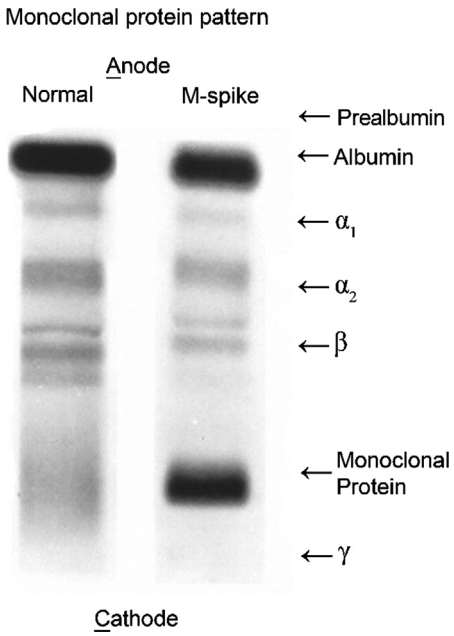


FIGURE 22.13 Control (normal) serum protein electrophoresis pattern *(left lane)* and serum protein electrophoresis showing a monoclonal gammopathy *(right lane)*.

Detection of an M-protein in serum or urine typically leads to further studies including the collection and examination of a bone marrow biopsy. The prevalence of various immunoglobulin subtypes in patients with MM is displayed in Table 22.5.

If MM and smoldering myeloma are ruled out and the M-protein is attributed to MGUS, then temporal

TABLE 22.7 Diagnostic laboratory focused criteria for immunoglobulin G and immunoglobulin A monoclonal gammopathies[a] including: monoclonal gammopathy of unknown significance,[b] smoldering myeloma, and symptomatic myeloma.

MGUS[c]	Smoldering multiple myeloma	Symptomatic myeloma
(all three criteria must be met)	**(both criteria must be met)**	**(all three criteria must be met)**
• M-protein in serum <3.0 g/dL • Bone marrow clonal plasma cells <10% and low level of plasma cell infiltration in a trephine biopsy (if done) • No related organ or tissue impairment (no end organ damage including bone lesions)	• M-protein in serum >3.0 g/dL and/or bone marrow clonal plasma cells >10% • No related organ or tissue impairment Absence of end-organ damage such as lytic bone lesions, anemia, hypercalcemia, or renal failure that can be attributed to a plasma cell proliferative disorder	• Clonal bone marrow plasma cells ≥ 10% • Presence of serum and/or urinary evidence of monoclonal protein (except in patients with true nonsecretory multiple myeloma) • Evidence of end-organ damage that can be attributed to the underlying plasma cell proliferative disorder, specifically: Hypercalcemia: serum calcium >11.5 mg/dL Renal insufficiency: serum creatinine >2 mg/dL or estimated creatinine clearance <40 mL/min Anemia: normochromic, normocytic with a hemoglobin value of >2 g/dL below the lower limit of normal or a hemoglobin value <10 g/dL Bone lesions: lytic lesions or severe osteopenia attributed to a plasma cell proliferative disorder or pathologic fractures

[a] *Occasionally patients with IgD and IgE monoclonal gammopathies have been described and are considered to be part of this category.*

[b] *Unless specifically distinguished, when the terms MGUS and multiple myeloma are used in general, they include IgM MGUS and light-chain multiple myeloma, respectively.*

[c] *Note that conventionally IgM MGUS is considered a subtype of MGUS.*

Adapted from reference International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma: Diagnosis, Staging & Monitoring, IMWG Publications. <http://imwg.myeloma.org/international-myeloma-working-group-consensus-criteria-for-response-and-minimal-residual-disease-assessment-in-multiple-myeloma/>, 2016 (accessed 01.12.18).

monitoring is clinically indicated. This is because a substantial proportion of patients with MGUS evolve into more serious conditions including MM. MGUS accounts for approximately 60% of unexpected discovery of M-proteins. The frequency of occurrence varies with age and is 1%–2% of individuals >50 years old, 3%–4% of individuals >70, and reportedly 10% in people >80 years of age. Patients diagnosed with MGUS require periodic monitoring with serum and urine studies, because the disease progresses in about 24% of these individuals [19], as shown in Fig. 22.14. Treatment of patients with MGUS is not recommended until laboratory or other abnormalities progress or symptoms of MM develop, because these patients may remain stable for years [16,17].

Multiple myeloma

MM is a malignancy that was first described in 1848 and is characterized by a proliferation of plasma cells of a single clone leading to a subsequent excess of monoclonal protein in serum (80% of cases) and/or urine (60% of cases). MM represents approximately 15% of all hematologic malignancies; it is the most common hematologic malignancy in African Americans and is the second most common in Caucasians. The lifetime risk of MM in the United States is about 0.76%. MM is a malignancy of plasma cells and may show clinical signs and symptoms including bone pain, pathologic fractures, weakness, malaise, hypercalcemia, spinal cord compression, renal failure, plasmacytoma, and neuropathies. Also, MM is characteristically associated with anemia. MM patients frequently have thrombocytopenia and leukopenia and impaired humoral immunity, which predisposes MM patients to infections, especially with encapsulated organisms such as *Pneumococcus*. The American Cancer Society estimates that about 30,770 new cases of MM in 2018 and that approximately 12,770 deaths were attributed to MM in that year. Table 22.7 shows the diagnostic criteria for symptomatic myeloma according to the IMWG [18].

For patients with symptomatic MM, chemotherapy is a first-line treatment. If the patient is classified as having asymptomatic MM, treatment is typically delayed until

TABLE 22.8 Diagnostic laboratory focused criteria for light-chain monoclonal gammopathies including light-chain monoclonal gammopathy of unknown significance.

Light-chain MGUS	Idiopathic Bence Jones proteinuria	Light-chain multiple myeloma[a]
(all criteria must be met)	**(all criteria must be met)**	**[the same as multiple myeloma (see Table 22.7) except no evidence of immunoglobulin heavy-chain expression]**
• Abnormal free light-chain ratio (<0.26 or >1.65) • Increased level of the appropriate involved light-chain (increased kappa free light chain in patients with ratio >1.65 and increased lambda free light chain in patients with ratio <0.26) • No immunoglobulin heavy-chain expression on immunofixation • Clonal bone marrow plasma cells <10% • Absence of end-organ damage such as CRAB that can be attributed to the plasma cell proliferative disorder	• Urinary monoclonal protein on urine protein electrophoresis ≥500 mg/24 h and/or clonal bone marrow plasma cells ≥10% • No immunoglobulin heavy-chain expression on immunofixation • Absence of end-organ damage such as CRAB that can be attributed to the plasma cell proliferative disorder	• Clonal bone marrow plasma cells ≥ 10% • Presence of serum and/or urinary monoclonal protein (except in patients with true nonsecretory multiple myeloma) • Evidence of end-organ damage that can be attributed to the underlying plasma cell proliferative disorder, specifically: Hypercalcemia: serum calcium >11.5 mg/dL Renal insufficiency: serum creatinine >2 mg/dL or estimated creatinine clearance <40 mL/min Anemia: normochromic and normocytic with a hemoglobin value of >2 g/dL below the lower limit of normal or a hemoglobin value <10 g/dL Bone lesions: lytic lesions or severe osteopenia attributed to a plasma cell proliferative disorder or pathologic fractures

CRAB, Hypercalcemia, renal insufficiency, anemia, and bone lesions; *MGUS*, monoclonal gammopathy of unknown significance.

[a]*Note that light-chain multiple myeloma is considered a subtype of multiple myeloma. Unless specifically distinguished, when the terms MGUS and multiple myeloma are used in general, they include IgM MGUS and light-chain multiple myeloma, respectively.*
Adapted from reference International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma: Diagnosis, Staging & Monitoring, IMWG Publications. <http://imwg.myeloma.org/international-myeloma-working-group-consensus-criteria-for-response-and-minimal-residual-disease-assessment-in-multiple-myeloma/>, 2016 (accessed 01.12.18).

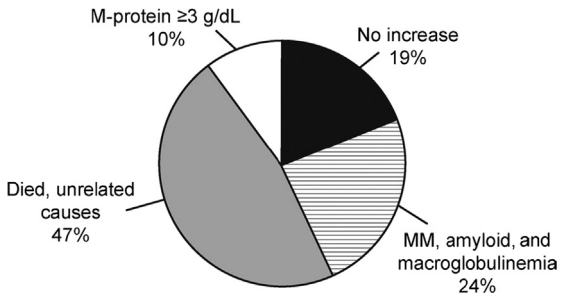


FIGURE 22.14 Findings at 20–35 years (median 22 years) in 241 patients initially presenting with monoclonal gammopathy of unknown significance. *Adapted from P.M. Ridker, E. Danielson, F.A.H. Fonseca, J. Genest, A.M. Gotto, J.J.P. Kastelein, et al., Rosuvastatin to prevent vascular events in men and women with elevated c-reactive protein,* New Engl. J. Med. ***359*** *(2008) 2195–2207.*

disease clinically progresses, for example, serum or urine concentrations of the M-protein substantially increase. Although no cure for MM is currently known, treatment results in improved patient outcomes including better survival, less pain, and fewer complications. The 5-year relative survival rate for MM is 46.6%, with younger patients showing better treatment response than the elderly [17]. Assessing treatment response is central to prognostication and management of MM patients. Table 22.9 displays the criteria indicating treatment response [18]. Serum FLC measurements, urine and serum agarose gel or capillary electrophoresis, quantification of the amount of protein comprising the M-protein in serum and urine, and IFE or immunosubtraction of serum and urine are all recommended measurements along with bone marrow examination and clinical criteria.

Laboratory testing is also important for temporal monitoring MM patients. Table 22.10 shows the criteria and laboratory support required for assessing disease progression or relapse according to current guidelines [18]. Serum FLC measurement, protein electrophoresis, IFE of serum and urine, and routine laboratory analysis play an important role in MM monitoring.

TABLE 22.9 International Myeloma Working Group uniform response criteria.

Response subcategory	Response criteria
Stringent complete response	Complete response as defined below plus • Normal free light-chain ratio • Absence of phenotypically aberrant plasma cells by multiparameter flow cytometry
Complete response	• Negative immunofixation on the serum and urine • Disappearance of any soft tissue plasmacytomas • <5% bone marrow plasma cells
Very good partial response	• Serum and urine M-protein detectable by immunofixation but not on electrophoresis OR • >90% reduction in serum M-protein plus reduction in 24-h urinary M-protein by >90% or to <100 mg/24 h
Partial response	• >50% reduction of serum M-protein and reduction in 24-h urinary M-protein by >90% or to <200 mg/24 h

Adapted from reference International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma: Diagnosis, Staging & Monitoring, IMWG Publications. <http://imwg.myeloma.org/international-myeloma-working-group-consensus-criteria-for-response-and-minimal-residual-disease-assessment-in-multiple-myeloma/>, 2016 (accessed 01.12.18).

TABLE 22.10 International Myeloma Working Group uniform response criteria: disease progression and relapse.

Relapse subcategory	Relapse criteria
Progressive disease	Requires at least one of the following • >25% increase in serum M-protein in 3 months (absolute increase must be >0.5 g/dL) • >25% increase in urine M-protein in 3 months (absolute increase must be >200 mg/24 h) • >25% increase in the difference between the involved and uninvolved free light-chain concentrations [applicable only to patients without measurable serum and urine M-protein (absolute increase must be >10.0 mg/dL)] • >25% increase in bone marrow plasma cell percentage (absolute percentage must be >10%) • Development of new bone lesions or soft tissue plasmacytoma • Development of hypercalcemia
Clinical relapse	Requires at least one of the following: • Development of new bone lesions or soft tissue plasmacytoma • Increase in size of existing plasmacytomas or bone lesions • Any of the following attributable to myeloma: Development of hypercalcemia Development of anemia (drop in Hemoglobin >2.0 g/dL) Rise in serum creatinine
Relapse from complete response	Requires at least one of the following • Reappearance of serum or urine M-protein by immunofixation or electrophoresis • Development of >5% plasma cells in the bone marrow • Appearance of any other sign of progression (e.g., new plasmacytoma and new lytic bone lesion)

Adapted from reference International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma: Diagnosis, Staging & Monitoring, IMWG Publications. <http://imwg.myeloma.org/international-myeloma-working-group-consensus-criteria-for-response-and-minimal-residual-disease-assessment-in-multiple-myeloma/>, 2016 (accessed 01.12.18).

Waldenström macroglobulinemia and immunoglobulin M multiple myeloma

IgM monoclonal gammopathies can present as several conditions including MGUS, smoldering Waldenström macroglobulinemia, and IgM MM. See Table 22.11 for criteria for their distinction [20]. Waldenström macroglobulinemia is the clinical expression of lymphoplasmacytic lymphoma, in contrast to the plasma cell proliferation in IgM MM that is associated with the secretion of high amounts of circulating monoclonal IgM protein. It is important to recognize that not all IgM

TABLE 22.11 Diagnostic laboratory focused criteria for Immunoglobulin M monoclonal gammopathies including smoldering Waldenström macroglobulinemia, Waldenström macroglobulinemia, and immunoglobulin M myeloma. For immunoglobulin M monoclonal gammopathy of unknown significance criteria, see Table 22.7

Smoldering Waldenström macroglobulinemia	Waldenström macroglobulinemia	IgM multiple myeloma
(Both criteria must be met)	**(All three criteria must be met)**	**(All three criteria must be met)**
• Serum IgM monoclonal protein ≥ 3 g/dL and/or bone marrow lymphoplasmacytic infiltration $\geq$ 10% • No evidence of anemia, constitutional symptoms, hyperviscosity, lymphadenopathy, or hepatosplenomegaly that can be attributed to the underlying lymphoproliferative disorder	• IgM monoclonal gammopathy (regardless of the size of the M-protein) • $\geq$10% bone marrow lymphoplasmacytic infiltration (usually intertrabecular) by small lymphocytes that exhibit plasmacytoid or plasma cell differentiation and a typical immunophenotype (e.g., surface IgM^+, $CD5^{+/-}$, $CD10^-$, $CD19^+$, $CD20^+$, and $CD23^-$) that satisfactorily excludes other lymphoproliferative disorders including chronic lymphocytic leukemia and mantle cell lymphoma • Evidence of anemia, constitutional symptoms, hyperviscosity, lymphadenopathy, or hepatosplenomegaly that can be attributed to the underlying lymphoproliferative disorder	• Symptomatic monoclonal plasma cell proliferative disorder characterized by a serum IgM monoclonal protein regardless of size • Presence of 10% plasma cells on bone marrow biopsy • Presence of lytic bone lesions related to the underlying plasma cell disorder and/or translocation *t*(11;14) on fluorescence in situ hybridization

IgM, Immunoglobulin M; *M-protein*, monoclonal protein.

Adapted from reference R.A. Kyle, "Benign" monoclonal gammopathy—after 20 to 35 years of follow-up,*Mayo Clin. Proc.* **68** (1993), 26–36..

monoclonal gammopathy is Waldenström marcroglobulinemia. In fact, the minority, only 15%, of IgM M-proteins are Waldenström macroglobulinemia, 55% are MGUS, and 25% are related to other lymphoproliferative disorders including chronic lymphocytic leukemia and other B-cell lymphomas. IgM MM only comprises 1%–2% of IgM monoclonal gammopathies.

The median age of onset for Waldenström macroglobulinemia is 63–68 years; clinical features include headache, dizziness, hepatomegaly, splenomegaly, anemia, and mental status changes; peripheral neuropathy occurs in approximately 5% of patients. Hyperviscosity, caused in large part by the high molecular weight of the IgM pentamer, is also a clinical feature and is related to the visual changes, retinal hemorrhage, and oral mucosal bleeding symptoms. Plasma viscosity must exceed fivefold that of water to produce symptoms, and values of ≥ 6 typically occur with plasma IgM concentrations >5 g/dL. Patients with plasma IgM concentrations of ≤ 4 g/dL are typically asymptomatic, so plasma viscosity measurements are not indicated. Waldenström macroglobulinemia differs from MM clinically in other important ways, including:

1. less bone pain and fewer lytic lesions;
2. fewer bacterial infections; and
3. greater incidence of serum hyperviscosity. Some textbooks state that virtually all patients have Bence Jones proteinuria. However, Fin et al. [7] reported that approximately 50% of patients excrete FLCs.

Cryoglobulinemia

Cryoglobulinemia is characterized by the presence of proteins that precipitate at temperatures cooler than normal body temperature (some at temperatures as high as 35°C). Cryoglobulins can be classified into three types.

1. Type I features an associated monoclonal protein (25% of all cases).
2. Type II mixes a monoclonal component and other proteins (25% of cases).
3. Type III is composed of mixed polyclonal proteins (50% of cases).

Patients with types II and III disease often have associated immune-complex disease.

Preanalytical factors are essential to detecting cryoproteins. Specimens from patients with suspected cyroglobulins must be maintained at 37°C throughout transportation to the laboratory and during specimen processing. If

temperatures are not appropriately maintained, cryoglobulins may precipitate out during the clotting process. After the sample clots, samples should be refrigerated at 4°C for 72 hours before assessing the presence of cryoprotein. Upon finding a cryoglobulin, the protein should be redissolved in saline warmed to 37°C, and then studied by, for example, electrophoresis and IFE for confirmation and characterization.

Serum from hepatitis C patients frequently demonstrates a type II cryoglobulinemia, which is typically IgM directed against polyclonal IgG. Hepatitis C may produce glomerulonephritis, leading to a combination of serum monoclonal proteins and renal failure that is not a consequence of light-chain deposition, light-chain casts, or amyloid.

Amyloidosis

AL amyloidosis, or "primary amyloidosis," is characterized clinically by peripheral neuropathy, carpal tunnel syndrome, congestive heart failure, nephrotic syndrome, enlargement of the tongue, and malabsorption. Most AL amyloidosis cases do not have evidence of plasma cell myeloma or lymphoma. However, most of these patients have a subtle B-cell disorder in which the cells produce the amyloidogenic light chains that are characteristic of the disease. The amyloid material consists of unbranched fibrils having a diameter of approximately 10 nm; 95% of the fibril proteins consist of λ light chains and 5% amyloid P-protein, arranged in a β-pleated sheet structure. The fibrils demonstrate distinctive staining with Congo red, which are salmon pink in color and show green birefringence under polarizing light. The amyloid material is deposited in blood vessels and also in the thin membranous mesangium of the kidney in a nodular or diffuse pattern.

Nonimmune forms of amyloidosis may be systemic and involve multiple organs. Unlike AL amyloidosis, some forms of true amyloid neither contain immunoglobulin components nor are associated with a monoclonal gammopathy. Therefore some patients with MGUS may also have an amyloid component that is coincidental and not causally related. Amyloid A (AA) amyloidosis, also termed "secondary AA amyloidosis," is a nonimmune form that is associated with chronic infection or inflammatory diseases such as rheumatoid arthritis, osteomyelitis, or granulomatous ileitis. Amyloid deposition in various tissues including the kidney results from the production of the amyloidogenic acute-phase protein serum AA during chronic inflammation. Appropriate treatment of the underlying chronic infection or inflammatory disease can slow down or even halt the progression. There are also familial forms of amyloidosis in which mutations of plasma proteins (e.g., TTR) are deposited in tissues. More than 50 different mutations in TTR are known in connection with this type of amyloidosis. Rarely, mutations also occur with other proteins including apolipoprotein Al, gelsolin, fibrinogen, and lysozyme and can cause familial amyloidosis.

Light-chain deposition disease

LCDD is associated with renal insufficiency and proteinuria or the nephrotic syndrome. This condition occurs when there is deposition of immunoglobulin light chains (typical ratio: $\kappa > \lambda$; 9:1) in the kidney along the tubular basement membrane of the distal tubules and loops of Henle. The thickened basement membrane shows characteristic diffuse immunofluoresence staining for λ and/or κ light chains. A large proportion (35%–40%) of LCDD patients meet criteria for MM. Microscopy on renal biopsy typically shows glomerular lesions that are described as nodular glomerulosclerosis with strongly periodic acid—Schiff-positive mesangial nodules; these lesions are similar to those observed in diabetic nodular glomerulosclerosis. Although LCDD is similar in some respects to amyloid, there is no "P" component, no staining with Congo red, and polarization shows no effect.

LCDD is a subset of conditions termed "monoclonal immunoglobulin deposition disease" that includes immunoglobulin light- and heavy-chain deposition disease (LHCDD) and heavy-chain deposition disease (HCDD). LHCDD mainly involves deposition of IgG-κ and IgG-λ components, whereas HCDD involves deposition of the γ heavy chain and no light-chain component.

Multiple sclerosis and cerebrospinal fluid electrophoresis

MS is a demyelinating disease that typically affects the brain, optic nerve, and spinal cord. Although advances in radiologic technology, such as MRI, have become increasingly important for diagnosis, 90%–95% of MS patients show oligoclonal banding on CSF agarose electrophoresis and IFE (Fig. 22.9). Oligoclonal banding is useful for diagnosis after other causes of banding such as CNS syphilis, chronic meningitis, and subacute sclerosing panencephalitis are excluded. The Q albumin is typically not increased in MS patients, because the disease does not often show increased permeability of the blood–CSF barrier. IgG index studies are less sensitive but are also considered useful for diagnosis and monitoring.

Because monoclonal proteins in serum may traverse the blood–CSF barrier and produce banding in CSF, performing electrophoresis on concomitant serum and CSF is important to avoid false-positives. IFE with IgG antisera allows more sensitive detection of oligoclonal banding. In addition to the conditions listed above, Guillain–Barré

syndrome, SLE, and bacterial meningitis may show bands in the CSF. Measurement of CSF FLC in the future may serve as an alternative to the labor-intensive IEF used for oligoclonal banding [21].

Summary

The analysis of proteins remains a critical component to laboratory medicine. Serum from patients being assessed for suspected MM and other cancers described herein is initially screened by SPEP or CZE. The interpretation of the initial SPEP or CZE pattern may lead to more sensitive diagnostic procedures such as IFE and serum FLCs. It is important that laboratorians understand the basic protein electrophoresis patterns and know the various exogenous and endogenous interferences that may affect patient's results. The introduction of m-tAb has also contributed to the overall success in the treatment and survival of these patients but presents unique challenges to the clinical laboratory. Current and future laboratorians will need to engage clinical partnerships in other areas of medicine to manage these new obstacles.

References

[1] P. Bobbert, A. Weithäuser, J. Andres, T. Bobbert, U. Kühl, H.P. Schultheiss, et al., Increased plasma retinol binding protein 4 levels in patients with inflammatory cardiomyopathy, Eur. J. Heart Fail. 11 (2009) 1163–1168.

[2] J.C. Won, C.Y. Park, S.W. Oh, S.W. Park, Increased plasma levels of retinol-binding protein 4 with visceral obesity is associated with cardiovascular risk factors, J. Diabetes Investig. 3 (2012) 457–463.

[3] M. van Hoek, A. Dehghan, M. Zillikens, A. Hofman, J. Witteman, E. Sijbrands, An RBP4 promoter polymorphism increases risk of type 2 diabetes, Diabetologia 51 (2008) 1423–1428.

[4] Nephrotic Syndrome. The Merck Manuals Online Medical Library. 2010. <www.merckmanuals.com> (accessed 30. 09.19).

[5] S.W. Cotten, Z. Shajani-Yi, M.A. Cervinski, T. Voorhees, S.A. Tuchman, N. Korpi-Steiner, Reference intervals and diagnostic ranges for serum free κ and free λ immunoglobulin light chains vary by instrument platform: implications for classification of patient results in a multi-center study, Clin. Biochem. 58 (2018) 100–107.

[6] B. Xu, Y. Tang, J. Zhou, P. Zhang, H. Li, Disease spectrum of abnormal serum free light chain ratio and its diagnostic significance, Oncotarget 8 (2017) 82268–82279.

[7] W.G. Finn, R. Gulbranson, S. Fisher, L. Rae Sample, R. Shalhoub, D. Hedstrom, et al., Detection of polyclonal increases in immunoglobulin g4 subclass by distinct patterns on capillary serum protein electrophoresis: diagnostic pitfalls and clinical observations in a study of 303 cases, Am. J. Clin. Pathol. 146 (303-11) (2016).

[8] P.M. Ridker, E. Danielson, F.A.H. Fonseca, J. Genest, A.M. Gotto, J.J.P. Kastelein, et al., Rosuvastatin to prevent vascular events in men and women with elevated c-reactive protein, New Engl. J. Med. 359 (2008) 2195–2207.

[9] G.M. Bernier, F.W. Putnam, Monomer—Dimer forms of bence jones proteins, Nature 200 (1963) 223–225.

[10] J.R. Genzen, D.L. Murray, G. Abel, Q.H. Meng, R.J. Baltaro, D. D. Rhoads, et al., Screening and diagnosis of monoclonal gammopathies: an international survey of laboratory practice, Arch. Pathol. Lab. Med. 142 (2018) 507–515.

[11] J.R. Mills, M.C. Kohlhagen, S. Dasari, et al., Comprehensive assessment of M-proteins using nanobody enrichment coupled to MALDI-TOF mass spectrometry, Clin. Chem. 62 (2016) 1334–1344.

[12] J.R. Mills, M.C. Kohlhagen, M.A.V. Willrich, T. Kourelis, A. Dispenzieri, D.L. Murray, A universal solution for eliminating false positives in myeloma due to therapeutic monoclonal antibody interference, Blood 132 (2018) 670–672.

[13] International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma: Diagnosis, Staging & Monitoring, IMWG Publications. <http://imwg.myeloma.org/international-myeloma-working-group-consensus-criteria-for-response-and-minimal-residual-disease-assessment-in-multiple-myeloma/>, 2016 (accessed 01.12.18).

[14] P.R. Greipp, J. San Miguel, B.G. Durie, et al., International staging system for multiple myeloma, J. Clin. Oncol. 23 (2005) 3412–3420.

[15] C.R. McCudden, J.F.M. Jacobs, D. Keren, H. Caillon, T. Dejoie, K. Andersen, Recognition and management of common, rare, and novel serum protein electrophoresis and immunofixation interferences, Clin. Biochem. 51 (2018) 72–79.

[16] S.V. Raijkumar, M.A. Dimopoulos, A. Palumbo, et al., International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma, Lancet Oncol. 15 (2014) e538–e548.

[17] C. Röllig, S. Knop, M. Bornhäuser, Multiple myeloma, Lancet 385 (2015) 2197–2208.

[18] International Myeloma Working Group Criteria for the diagnosis of multiple myeloma. Guideline, IMWG Publications. <http://imwg.myeloma.org/international-myeloma-working-group-imwg-criteria-for-the-diagnosis-of-multiple-myeloma/>, 2015 (accessed 01.12.18).

[19] R.A. Kyle, "Benign" monoclonal gammopathy—after 20 to 35 years of follow-up, Mayo Clin. Proc. 68 (1993) 26–36.

[20] S.V. Rajkumar, Multiple myeloma: 2014 update on diagnosis, risk-stratification, and management, Am. J. Hematol. 89 (2014) 999–1009.

[21] K.M. Gurtner, E. Shosha, S.C. Bryant, B.D. Andreguetto, D.L. Murray, S.J. Pittock, et al., CSF free light chain identification of demyelinating disease: comparison with oligoclonal banding and other CSF indexes, Clin. Chem. Lab. Med. 56 (2018) 1071–1080.

Suggested readings

[22] I.T. Whicher, C.P. Price, K. Spencer, Immunonephelometric and immunoturbidimetric assays for proteins, Crit. Rev. Clin. Lab. Sci. 18 (1983) 213–260.

[23] K. Felgenhauer, Differentiation of the humoral immune response in inflammatory diseases of the central nervous system, J. Neurol. 228 (1982) 223–237.

[24] R.A. Kyle, "Benign" monoclonal gammopathy, JAMA 251 (1984) 1849–1854.
[25] D. Thompson, A. Milford-Ward, J.T. Whicher, The value of acute phase proteins measurements in clinical practice, Ann. Clin. Biochem. 29 (1992) 123–131.
[26] R.A. Kyle, Newer approaches to the management of multiple myeloma, Cancer 72 (1993) 3489–3494.
[27] E. Rouah, B.B. Rogers, G.J. Buffone, Transferrin analysis by immunofixation as an aid in the diagnosis of cerebrospinal fluid otorrhea, Arch. Pathol. Lab. Med. 111 (1987) 756–757.
[28] A.M. Johnson, E.M. Rohlfs, L.M. Silverman, Proteins, in: C.A. Burtis, E.R. Ashwood (Eds.), Tietz Textbook of Clinical Chemistry, third ed., WB Saunders, Philadelphia, PA, 1999, pp. 477–540.
[29] C.M. Schambeck, R. Bart, W. Hochtlen-Vollmar, M. Wick, R. Lamerz, A. Fateh-Moghadam, Characterization of myeloma cells by means of labeling index, bone marrow histology, and serum f32microglobulin, Am. J. Clin. Pathol. 106 (1996) 64–68.
[30] R.A. Kyle, S.V. Rajkumar, Monoclonal gammopathies of undetermined significance, Rev. Clin. Exp. Hematol. 2 (2002) 225–232.
[31] R.A. Kyle, J.P. Garton, The spectrum of IgM monoclonal gammopathy in 430 cases, Mayo Clin. Proc. 62 (1987) 719–731.

Self-assessment questions

1. Prealbumin is which of the following?
   a. A protein rich in nonessential amino acids
   b. A protein that migrates closer to the cathode than ALB
   c. The precursor protein for ALB
   d. The same protein as TTR
   e. The major plasma binding protein for copper
2. Which of the following proteins is a protease inhibitor?
   a. C-reactive protein
   b. A1AT
   c. Transferrin
   d. Haptoglobin
   e. Fibrinogen
3. A band is visualized in the γ on SPEP. However, on IFE, no band that reacts with IgG, IgA, IgM, or κ or λ antisera is identified. Which of the following is a possible cause?
   a. The band is possibly IgD.
   b. The specimen may be heparinized plasma, so the band could be fibrinogen.
   c. The band could be transferrin.
   d. a–c are possible
4. Which of the following is a difference between nephelometry and turbidimetry measurements?
   a. The immunoprecipitin curve is not relevant to nephelometry measurements, whereas turbidimetry depends on this curve.
   b. Turbidimetry is based on Rayleigh scattering, whereas nephelometry is not.
   c. Detection of nephelometry signals is typically made at a 30% offset angle from the incident light source, whereas turbidimetry is at a 180-degree angle.
   d. Turbidimetry is more difficult to adapt to spectrophotometric equipment compared with nephelometry.
   e. All of the above are differences between nephelometry and turbidimetry.
5. Which is a calculation that can be used for assessing the integrity of the blood–CSF barrier?
   a. IgG index
   b. Globulin calculation
   c. Q albumin
   d. TIBC
   e. IgG synthesis rate
6. Which of the following is a correct difference between Waldenström macroglobulinemia and multiple myeloma?
   a. More visual disturbances occur with Waldenström macroglobulinemia.
   b. More bone pain and more lytic lesions occur with Waldenström macroglobulinemia.
   c. A greater incidence of serum hyperviscosity occurs with Waldenström macroglobulinemia.
   d. a and c are correct differences between Waldenström macroglobulinemia and multiple myeloma.
   e. a–c are correct differences between Waldenström macroglobulinemia and multiple myeloma.
7. Which of the following is a diagnostic criterion for asymptomatic (smoldering) myeloma?
   a. $<10\%$ plasma cells in bone marrow aspirate
   b. No evidence of lytic bone lesions
   c. No reduction in uninvolved serum immunoglobulins
   d. No light chains excreted in urine
8. Which of the following is correct about amyloidosis?
   a. The majority of AL amyloidosis cases do not have evidence of plasma cell myeloma or lymphoma.
   b. Immunoglobulin components are a characteristic of all forms of amyloidosis.
   c. The fibrils in AL amyloidosis demonstrate distinctive staining with Congo red that is salmon pink in color, but no color is observed under polarizing light.
   d. AA amyloidosis is also termed "primary amyloidosis."
   e. Serum AA deposition is the principle cause of renal injury in AL amyloidosis patients.
9. Which of the following is a negative acute-phase protein?
   a. A1AT
   b. Haptoglobin
   c. Ceruloplasmin
   d. Orosomucoid
   e. None of the above

10. Measurement of FLC is an addition to laboratory medicine that aids in the evaluation and management of multiple myeloma and related clonal plasma cell disorders. Which of the following makes FLC assay different from and advantageous over the electrophoretic assays?
    a. FLC assay identifies the light-chain type of the monoclonal protein more effectively than electrophoresis.
    b. FLC assay can be used for screening, prognosis, monitoring for recurrence, and monitoring therapeutic response, all more effectively than electrophoresis.
    c. FLC assay can be used in the quantitative monitoring of nearly two-thirds of patients who had previously been deemed to have nonsecretory myeloma.
    d. FLC assay negates the need of immunofixation study on urine specimen.
    e. All of the above.
11. Which of the following medications may appear on conventional agarose SPEP?
    a. Sulfamethoxazole
    b. 5-Fluorocytosine
    c. Piperacillin
    d. Siltuximab
    e. Hydroxycobalamin
12. Which of the following type of cyroglobulins may be seen in serum of patients with Hepatitis C?
    a. Type I
    b. Type II
    c. Type III
    d. Type IV
    e. Type V

Answers

1. d
2. b
3. b
4. c
5. c
6. d
7. b
8. a
9. e
10. c
11. d
12. b

Chapter 23

The complement system

Maria Alice V. Willrich

Department of Laboratory Medicine and Pathology, Division of Clinical Biochemistry and Immunology, Mayo Clinic, Rochester, MN, United States

Learning objectives

1. Define the different pathways for activation of the complement cascades.
2. Differentiate the assays used for assessment of complement function and quantitation, as well as assays used for the measurement of autoantibodies and activation fragments of complement.
3. Apply the various test combinations for aiding in diagnosis of complement disorders or monitoring complement consumption.
4. Classify the disorders associated with complement abnormalities as acquired or genetic in terms of underlying pathology.
5. Propose tests utilization approaches for the various complement disorders.

Introduction

The complement system is part of our innate immune response and consists of a number of precursor cell-bound or soluble proteins. To date, over 50 circulating proteins, enzymes, proenzymes, cofactors, and control proteins have been described, with an additional 15 cell-associated proteins involved in signaling, clearance, and phagocytosis. The complement system was discovered in 1891 and named alexin. In 1899 it was renamed complement, given its role in complementing the functions of the cellular immune system and antibodies. The complement system is responsible for infection and inflammation control by directly killing bacterial pathogens via cell lysis, or tagging infected cells or pathogens with opsonins for downstream removal by phagocytes. Anaphylatoxins released with complement activation will promote inflammation and increase vascular permeability for innate immune cell passage. In addition, immune complex clearance and tagging of cell debris and DNA fragments are some of the complement system's main housekeeping roles.

With the advent of eculizumab, a complement inhibitor that targets component C5, proper complement testing for diagnoses, and therapeutic monitoring are warranted. Outside the study of rare immunodeficiencies, which may have genetic and acquired causes, the complement system has been recognized to be involved in many conditions beyond infection. Complement testing should be considered when there is suspicion of primary immunodeficiencies presenting in the form of angioedema or recurrent pyogenic infections, or family history of complement deficiencies. Autoimmune diseases such as systemic lupus erythematosus (SLE), a disease in which increased amounts of immune complexes are generated, are often associated with hypocomplementemia. Thus testing patients throughout the SLE disease course is common. Testing is also indicated when there is tissue injury that can potentially be mediated by the complement system, such as graft rejection, organ damage observed in kidney glomerulopathies, or intravascular and extravascular hemolysis present in conditions such as hemolytic uremic syndrome (HUS) or paroxysmal nocturnal hemoglobinuria (PNH). However, with so many circulating and cell-associated proteins, understanding which tests are important is challenging and requires further understanding of the complement system to optimize test utilization and interpretation.

Overview of the complement system

In physiological conditions, complement is the first line of defense and surveillance of the immune system. Complement provides a mechanism for continuous immune surveillance through a network of soluble and bound proteins that work together to achieve molecular pattern-based immune recognition, thereby triggering a complement response when microbial pathogens, foreign antigens, or danger-associated molecular patterns (DAMPs) are recognized. The complement system circulates pattern-recognizing molecules (PRMs), including C1q, mannose-binding lectin (MBL), ficolins, and properdin, which are able to recognize pathogen-associated molecular patterns (PAMPs) or DAMPs for autoactivation and initiation of a serine protease cascade. This cascade,

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00023-5

also known as the complement cascade, results in opsonization and potential lysis of the target microbe [1–4].

Complement components are secreted primarily by the liver. Exceptions are C1q (secreted by brain microglia, immature dendritic cells, monocytes, and macrophages), C7 (monocytes, macrophages, endothelial cells, fibroblasts, and central nervous system cells), properdin (mainly produced by neutrophils), and factor D (FD, synthesized by adipocytes). There are three main pathways for complement activation, and these are the classical pathway (CP), the lectin pathway (LP), and the alternative pathway (AP) (Fig. 23.1). The CP is activated by an antigen–antibody immune complex or C-reactive protein (CRP), whereas the LP is activated directly by PAMPs such as mannose-containing bacterial surfaces. The AP is constitutively active at low levels and can be activated by spontaneous hydrolysis of C3 when a potential threat is detected. All three pathways are unique in the types of PRMs that activate the complement cascade; however, all lead to formation of an amplification loop mediated by enzymatic complexes called convertases. The C3 convertase is able to initiate opsonization of complement targets. Opsonization is the process of tagging molecules for removal or phagocytosis with antibodies (e.g., IgG) or complement fragments. Opsonization more commonly involves the C3b fragment, which is generated after C3 convertase-mediated C3 cleavage. Activated C3 convertase subsequently favors formation of a C5 convertase that will initiate cell lysis via formation of a membrane attack complex (MAC). The anaphylatoxins C3a and C5a have potent inflammatory effects and promote chemotaxis. They are released after C3 and C5 cleavage, respectively. Complement activation is controlled by a set of membrane-bound and fluid phase regulators to prevent overactivation.

The three complement activation pathway cascades are described in detail in the next sections.

The classical pathway

The complement CP is activated through the PRM C1q [5]. Endogenous C1q activators include IgM and IgG immune complexes, surface-bound pentraxins (such as CRP), β-amyloids fibrils, and tissue-damaged elements from apoptotic cells [6–8]. Exogenous C1q activators consist mostly of molecular patterns on pathogen surfaces such as bacterial porins [2]. C1q binds to two C1r and two C1s serine proteases to form a $C1qC1r_2s_2$ pentamer. The C1 complex in circulation is able to recognize PAMP or DAMP activators through C1q. When the target is bound to C1q, it initiates a conformational change in the complex, allowing C1r to autocleave in order to self-activate and, in turn, activate C1s (Fig. 23.2). Activated C1s subsequently cleaves C4 to generate C4a (released to circulation) and C4b (opsonizing) fragments. C4 cleavage by surface-bound C1 complex will result in activators being tagged with C4b. The surface-bound C4b recruits C2, which is cleaved by the C1 complex to release a C2b fragment (to circulation) and a C2a serine protease fragment; the latter remains associated with C4b.

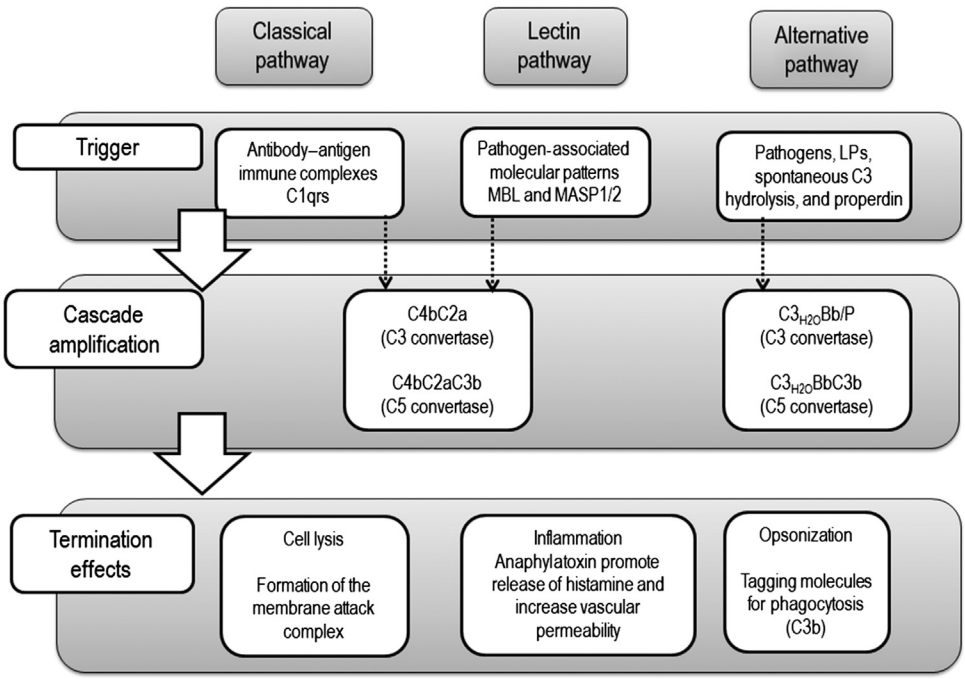


FIGURE 23.1 Overview of complement system. The complement pathways have unique pattern-recognizing molecules that recognize the different threats and trigger the cascade activation. All pathways converge to the formation of potent convertase complexes that target C3 and C5. The complement cascades effects include opsonization, inflammatory responses, and cell lysis via formation of the MAC.

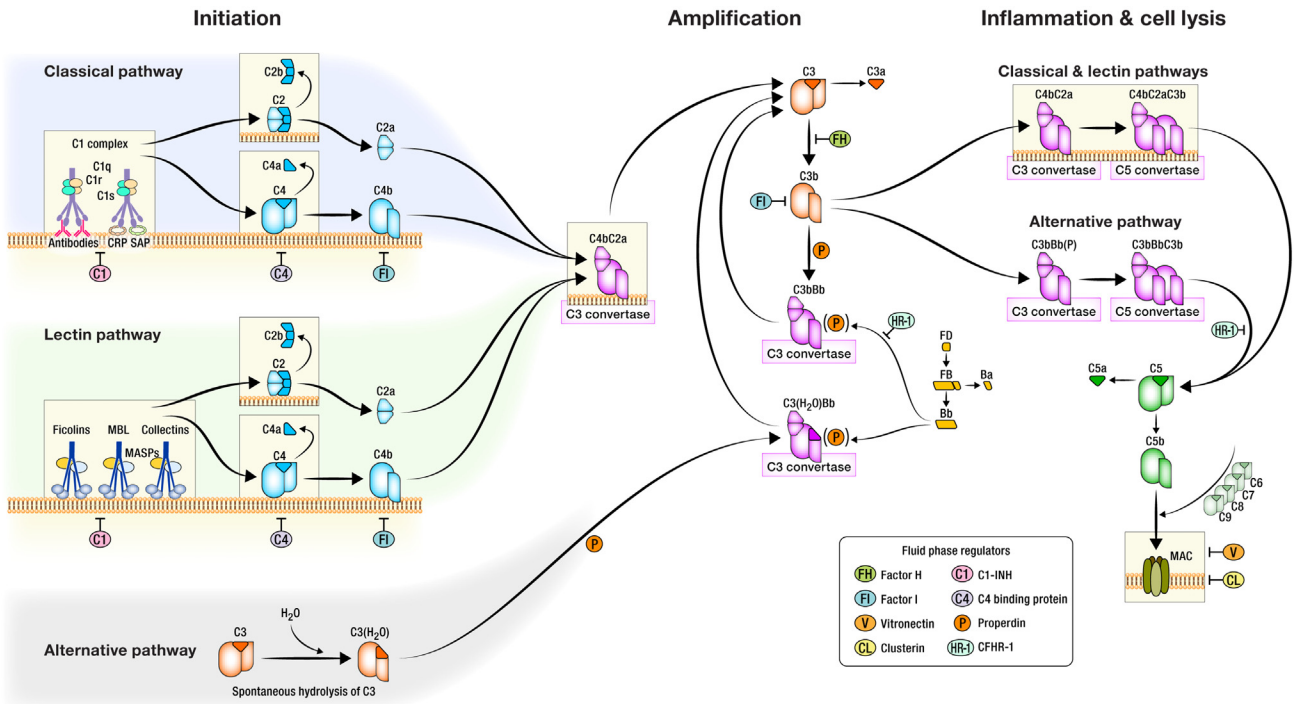


FIGURE 23.2 The complement cascades. Complement initiation, amplification, regulation, and formation of the terminal complement complex. *Initiation, top*: Classical pathway initiation via C1q and the C1 complex to form the C3 convertase (C4bC2a). *Initiation, middle*: Lectin pathway initiation via ficolins, mannose-binding lectin, and collectins to form the C3 convertase (C4bC2a). The roles of C1 inhibitor, C4BP, and factor I in regulating classical pathway and lectin pathway initiation are highlighted. *Initiation, bottom*: Alternative pathway initiation via the "tick-over" mechanism (spontaneous hydrolysis of C3) to form the alternative pathway C3 convertase $C3_{H2O}Bb$. *Amplification*: The amplification loop shows the role of properdin in complement initiation and amplification. The roles of factor I, factor H, and CFHR1 (HR-1) in regulation of complement amplification are highlighted along with the role of factor D and factor B in complement cascade amplification. *Inflammation and cell lysis*: The role of the amplification loop in generating the C5 convertases is shown along with the C5 convertase formation of the terminal complement complex. *Adapted from A. Frazer-Abel, L. Sepiashvili, M.M. Mbughuni, M.A. Willrich, Overview of laboratory testing and clinical presentations of complement deficiencies and dysregulation. Adv. Clin. Chem. 77 (2016) 1–75 [9] with permission of the Mayo Clinic Foundation, retainer of the copyrights for publication.*

The formation of the surface-bound C4bC2a protease complex is also known as C3 convertase. The C3 convertase plays a central role in complement signal cascade amplification and activation of the terminal pathway (TP). First, it cleaves C3 into C3a and C3b. C3a is an anaphylatoxin, and its generation starts a series of proinflammatory events. Similar to C4b, C3b is an opsonin, a molecule that tags complement targets for opsonization, facilitating phagocytosis and removal. It also acts as a surface for complement cascade amplification through factor B (FB) and FD from the AP (Fig. 23.2). Since all three complement activation pathways converge by assembling a C3 convertase, which has a high affinity for C3, C3b serves as the major opsonizing agent of the complement system.

The lectin pathway

Collectins and ficolins are the PRMs that activate the LP. They recognize membrane glycoprotein and acetylated glycolipid patterns on pathogen membrane surfaces. Once bound to activating surfaces, collectins (MBL or collectin LK) or ficolins recruit mainly MBL-associated serine protease 1 (MASP1) and MBL-associated serine protease 2 (MASP2), which are proteases structurally similar to the C1r and C1s enzymes from the CP. MASP1 binding to target surfaces results in autoactivation and subsequent activation of MASP2, which in turn will trigger cleavage of C4 and C2 to yield C4a, C4b, C2a, and C2b fragments. As in the CP, C2b and C4a fragments are released into circulation; whereas the C4b opsonin is capable of recruiting C2a to form a surface-bound LP C3 convertase identical to the CP C3 convertase, but containing C4bC2a [4].

The alternative pathway

The AP is the dominant complement surveillance method using a tick-over mechanism for complement activation, where spontaneous hydrolysis of a buried thioester bond in C3 releases a $C3_{H2O}$ fragment. This fragment has C3b-like properties. $C3_{H2O}$ binds to FB to form a proconvertase complex, $C3_{H2O}FB$ (Fig. 23.2). FB in the proconvertase complex is then cleaved by FD to release an

inactive Ba fragment and a Bb serine protease fragment, which associates with $C3_{H2O}$ to yield a soluble AP C3 convertase, $C3_{H2O}Bb$. The soluble C3 convertase is similar to the CP or LP C3 convertases in that it also cleaves C3 to generate a C3b fragment, but in this case, the fluid phase convertase makes a C3b fragment that must diffuse in search of complement-activating PAMP or DAMP targets or any reactive or unprotected surface. C3b generated from CP and LP C3 convertases is also able to recruit FB and FD to form the AP surface-bound C3 convertase (C3bBb), facilitating cross talk of CP, LP, and AP during complement cascade amplification.

In the AP, the PRM properdin is the only positive regulator of complement and has the capability to adhere C3bBb onto activating cell surfaces, forming a C3bBbP convertase (Fig. 23.2). Properdin can also bind to soluble C3b, $C3_{H2O}$, $C3_{H2O}Bb$, and C3bBbP, thereby substantially increasing their liquid phase stability and the likelihood of reacting with a nearby surface where assembly of a C3 convertase can occur. Target-bound properdin can recruit circulating C3b, $C3_{H2O}$, $C3_{H2O}Bb$, or C3bBb components to surfaces where these AP cascade proteases are stabilized. In the end, the AP omits components C1, C2, and C4, and is thought to be the dominant mechanism for complement amplification through C3, FB, FD, and properdin. Due to its ability to act as an amplification loop, the AP is often credited with nearly 80%−90% of total complement activation through FB and FD, even when the initial signal comes from C3b fragments produced by a C3 convertase from the CP or the LP.

The terminal pathway

Once the complement system is activated and C3 convertase is abundant, the TP is activated when C3b is high enough that CP, LP, or AP C3 convertases are able to bind additional C3b fragments, shifting the substrate specificity of the convertases from the C3 to the C5 component. In principle, excess C3b converts C3 convertases to C5 convertases [1]. C5 convertases then cleave C5 to yield C5a and the C5b fragments. C5a is the most potent anaphylatoxin generated by the complement system. C5b associates with C6 and C7, forming a soluble or surface-bound C5b-7 complex able to bind C8 and multiple C9 proteins to yield a soluble or surface-bound MAC (Fig. 23.2). The MAC in circulation is also known as the terminal complement complex, and can be held in the soluble form by the binding of vitronectin or clusterin yielding sC5b-9, and controlling the lytic properties of the MAC.

Complement regulation

With such a powerful system directed to lysis and opsonization, it is important that complement also has mechanisms for host cell protection. Complement regulation occurs through a number of soluble and membrane-bound proteins able to suppress complement activation at multiple checkpoints. Key molecules involved in complement regulation include C1 inhibitor (C1-INH), factor H (FH), complement factor H-like protein 1 (FHL-1), factor I (FI), factor H-related proteins (CFHRs), C4-binding protein (C4BP), CD46 [membrane cofactor protein (MCP)], CD55 or decay-accelerating factor, and carboxypeptidase-N.

Nonspecific CP and LP activation is prevented by the acute phase reactant C1-INH, which performs direct inhibition of the C1r/C1s or MASP1/MASP2 proteases in the CP or LP, respectively, to cause irreversible inactivation of these serine proteases in the absence of appropriate PRMs for complement activation with covalent binding and subsequent removal from circulation. When C1r/C1s and MASP1/MASP2 proteases are activated, regulatory checkpoints then shift to regulation of C3 and C5 convertase activities. At this point, key complement regulators include FH, FI, C4BP, CD55, and CD46 (Fig. 23.2). FI regulates all C3 and C5 convertases from all pathways through proteolytic inactivation of C3b and C4b, where FI is the direct inactivating enzyme. FI has a number of cofactors that direct its action. CD55 is a cell-bound regulator that facilitates accelerated decay of complement convertases through competitive binding of C3b and C4b. More tailored regulation of CP and LP convertases also occurs through C4BP (Fig. 23.2), with the ability to speed decay of CP/LP C3 convertase by binding to C4b to cause displacement of C2a or by acting as a cofactor for FI to aid in C4b proteolytic inactivation.

FI and FH are essential inhibitors of the AP C3 and C5 convertases, which also occur in order to keep in-check complement activation through the tick-over mechanism.

FH is a humoral complement regulator that plays an important role in distinguishing self from nonself. By binding to host cell surface glycans such as glycosaminoglycans (GAGs), FH localizes onto self surfaces where it protects host cells from complement by inhibition of AP C3 and C5 convertases. FH inhibits AP C3 and C5 convertase by acting as a cofactor for FI and also through competitive binding to C3b and FB in order to accelerate AP convertase decay. FH has high homology with many related proteins with similar or antagonistic roles in complement regulation. FHL-1 is an alternatively spliced version of FH where the C-terminus is truncated. Like FH, FHL-1 also recognizes self-specific molecular patterns, with a decay-accelerating activity for convertases and a cofactor role for FI to suppress complement activation on host surfaces. Complement factor H-related (CFHR) proteins 1 through 5 are proteins encoded by separate genes with some homology to FH. These proteins have the

ability to bind C3b and distinguish between self and nonself surfaces. The role of CFHR proteins in complement regulation is poorly understood and ranges from complement activation through competition with FH binding on host cell surfaces to complement regulation through inhibition of C3 and C5 convertases. CFHR-1 can bind to C3b in the C5 convertase and inhibit C5 cleavage, thereby inhibiting MAC formation. CFHR-1 also inhibits C3 convertase by inhibiting Bb binding to C3b.

Finally, carboxypeptidase-N is a protease able to cleave and inactivate the complement anaphylatoxin peptides C3a and C5a. While removal of the terminal desArg from C3a by carboxypeptidase-N eliminates the anaphylatoxin activity of C3a, C5a retains some of its activity after cleavage [10]. CD59 is glycophosphatidylinositol (GPI)-anchored protein expressed by host cells and functions to inhibit formation of the MAC on host cell surfaces [11]. Another important complement regulator is vitronectin, which regulates MAC formation by binding to C7.

Assays in the clinical laboratory

Serologic complement assays

Measurement of the complement system proteins in serum or plasma can be divided into four main categories:

1. measurement of complement function or lytic activity,
2. quantitation of complement components,
3. detection of autoantibodies against complement components, and
4. quantitation of activation fragments or split products.

CH50 and AH50 are the most common functional assays to measure CP and AP functions, respectively. LP function assays are also available, although not as common in the clinical setting as the assays for CP and AP given the more limited clinical utility. These assays are performed in serum samples and require the complement cascade to be activated in vitro; the reactions progress until the formation of a measurable MAC occurs. In a few assay formats, antibodies targeting the neoepitope of C5b-9 are used, while in others, the lytic effect of MAC is measured. Reports are usually in % activity or arbitrary units.

For quantitation of complement components, there are a number of automated and manual methods available, which measure the amount of antigen in the sample, most commonly reported in mass units such as milligrams per milliliter or micrograms per deciliter. Detection of autoantibodies to complement components will aid in differentiating genetic from acquired conditions. Finally, the quantification of activation products gives an estimation of the activation state of the pathways and determines if a complement protein is reduced due to either increased consumption or decreased synthesis. The complement cascades are dependent on calcium and hence using citrate or EDTA as chelating agents prevents in vitro complement activation. Serum is the required specimen when functional assays are carried out, and plasma samples are recommended when studying the activation fragment concentrations.

Preanalytical considerations

As the heat labile fraction of the immune system, complement activation in vitro has been the main challenge to overcome by researchers and scientists. Complement components of the CP are activated by immunoglobulins or immune complexes, and the AP can be activated by spontaneous hydrolysis of C3. Even the clotting process has been known to activate complement cascades; therefore, for activation fragment assays, plasma is the preferred sample source. Both degradation by temperature and consumption of complement components after in vitro activation will lead to falsely low function measurements and falsely high concentrations of activation fragments, respectively, thereby complicating the assessment of complement dysregulation. While preanalytic handling can lead to falsely abnormal results, it is far less likely that such issues could lead to falsely normal results.

If more than one component is measured as low, it is important to look for poor sample handling, use of IV catheters or other underlying disease states that could be causing chronic complement activation. A possible complement autoantibody, deficiency in a control protein, malnutrition, protein-wasting states, or age (newborn state) may also contribute to abnormal results. An additional consideration is that complement assays may be ordered by physicians in several conditions where standard treatment includes plasmapheresis or plasma exchange. The procedure itself, if traumatic, may activate complement, and the results will not necessarily reflect what is going on with the patient's complement system. Thus it is recommended to collect blood prior to plasma exchange whenever possible. Abnormal complement functional tests or low complement concentrations inconsistent with the clinical history should be verified with a new blood draw. It is recommended to freeze specimens immediately postcollection. Long-term stability is optimal when the sample is kept at $-70^{\circ}C$ prior to testing [12].

Postanalytical challenges

In addition to the preanalytical challenges associated with complement lability, proficiency testing is another area that suffers from complement poor stability. Most analytes are not regulated by proficiency testing agencies, such as the College of American Pathologists or the Food and Drug Administration (FDA), and a range of assays

are available as laboratory developed tests. Commercial proficiency testing is primarily available only for the major laboratory tests CH50, C3, and C4. Thus alternative assessment of performance with split blind samples is a common practice to address the lack of commercial proficiency testing. However, this approach has contributed to the lack of standardization or harmonization in reporting practices. To address this, the International Complement Society (www.complement.org), under the International Union or Immunological Societies, has sponsored a proficiency testing program under INSTAND (Society for Promoting Quality Assurance in Medical Laboratories e. V.). Since many assays are neither standardized nor harmonized, which means the reference intervals and performance between methods are often not interchangeable across laboratories, the assays are evaluated on a qualitative or semiquantitative scale for agreement and consensus. Ideally, individuals who require serial testing should be monitored within the same laboratory, and when crossing healthcare institutions, the comparison of test results should be minimal or reviewed in the qualitative or semiquantitative scale, with understanding of the differences between methodologies and reference intervals.

Given the complexity of testing and relationship to so many different clinical diseases, several tests are needed for a comprehensive overview of complement function. A panel approach is usually required to isolate a complement disorder within the CP, LP, or AP. Examples of such panels can be found from a number of clinical laboratories that specialize in complement or specific complement-related diseases.

Methods for analysis of complement function or activity

The study of complement system function may be performed using manual radial immunodiffusion (RID) hemolytic assays, or using automated methods such as liposomal lysis and ELISAs. Hemolytic functional assays are based on the ability of complement to induce red blood cell (RBC) lysis after formation of an MAC; these assays shape the initial approach to complement testing.

RID is a simple and versatile method, without the need for sophisticated instrumentation, and requires a small sample volume for testing. RID employs an agarose gel containing antiserum to a given complement component, when the goal is to assess complement component concentrations. Alternatively, RID agarose gels may contain heterologous RBCs from sheep, rabbit, or chicken for functional hemolytic assays. For both forms of the assay, wells are punched onto the gel, and samples are applied and allowed to diffuse passively for a given amount of time, usually in the scale of 12−36 hours. On an RBC agarose gel, hemolysis will be observed around the samples' wells. The diameter around the well is proportional to complement function. RID may be affected by variations in temperature, humidity, and incubation time. The manual nature of the technique makes it labor intensive, and incubation periods range from hours to days.

For CH50 functional testing, sheep RBCs are coated with antirabbit IgM antibodies, with or without IgG. The animal RBCs coated with IgM will trigger the activation of the CP. The dilution of patient's serum needed to lyse 50% of RBCs is then determined, hence the origin of the test name. The assay requires the presence and functionality of all proteins of the CP and TP. Deficiency in C1 or C2−C8 will result in no-to-little lysis of RBCs, with observed hemolysis of less than 5%. C9-deficient patients may have residual CH50 function, with less than 30% of RBCs hemolyzed. A low CH50 result may occur due to congenital complement deficiencies, increased consumption of complement components evidenced by infection or acute flares in autoimmune diseases, such as SLE, with abundant immune-complex formation, or insufficient synthesis of complement components (e.g., hepatic diseases). In the AP function assay, or AH50, a buffer is used to block the activation of the CP and LP (Mg EGTA is used to chelate calcium, necessary for the activity of CP and LP, but not AP). Rabbit or guinea pig RBCs are then used to spontaneously activate the AP. The assay requires all components in the AP and the TP. Patients with defects in the regulatory proteins FH, FI, or in the components FB and FD have a constitutively active AP, resulting in consumption of the early components of the pathway. When laboratory testing is performed during the acute phase of the disease, AH50 activity is often low. Little or no lysis is observed for deficiencies of C3−C9, FD, and FB.

As a second line of testing, the functionality of individual complement components may be tested by mixing a patient's serum with complement-depleted serum using similar approaches as CH50 and AH50 assays. If CH50 is decreased and AH50 is within reference intervals, a physician may try to identify the CP-deficient component: C1, C2, or C4, and specific functional tests may be ordered for each component. If hemolysis occurs after mixing a patient's serum with complement C2-depleted serum, for instance, it becomes evident that the C2 is present in the patient's serum and functional. If, on the other hand, the mixing study did not *complement* the deficient serum and there is still little or no hemolysis, this is evidence that the C2 is missing or nonfunctional in the specimen of interest.

ELISAs may be designed as functional assays that measure neoepitopes generated after complement activation. Fig. 23.3 shows conventional ELISA methods for measurement of the activity within the CP, LP, and AP. Three ELISA systems are illustrated. For the CP, IgM is

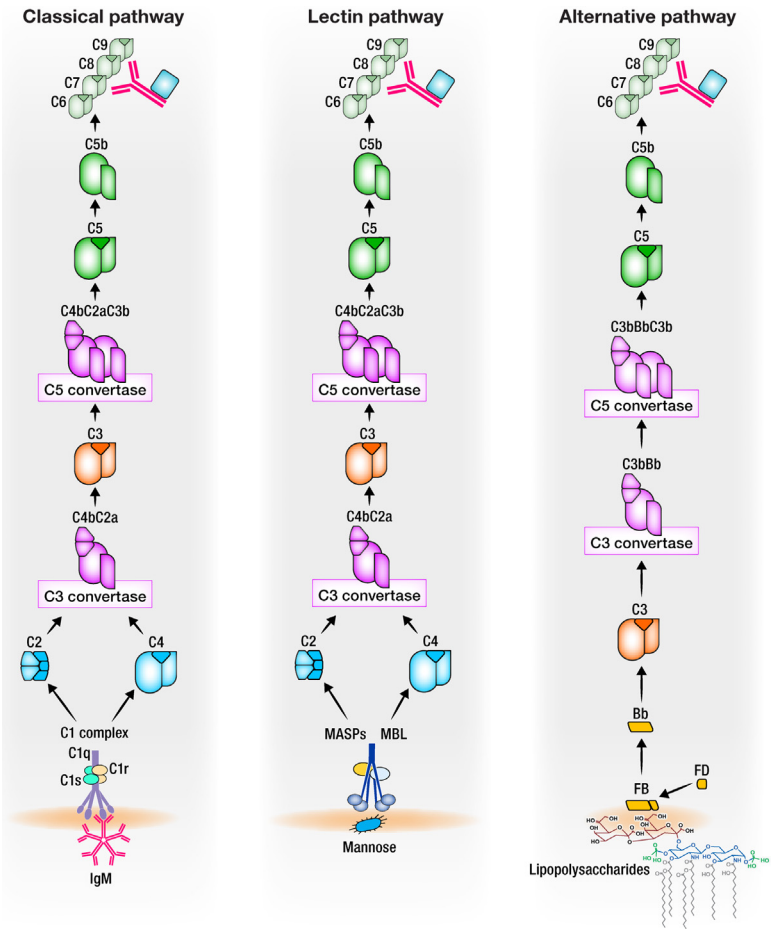


FIGURE 23.3 ELISA assays are available to measure function of classical, lectin, and alternative pathways of complement as described in [13].

coated on the solid-support ELISA plate. For the LP, mannan is on the plate, and, for the AP, lipopolysaccharides are from Gram-negative bacteria. The CP and LP assays use diluted sera 1:101, while the AP assay uses a 1:18 dilution. The LP function assay also uses antibodies to C1q to block the CP, and AP is minimized using the higher dilution schemes. These characteristics make the assays pathway-specific. In all of the tests, the complement system in a patient's serum will be activated, the respective cascades amplified to create C3 and C5 convertases, and ultimately the MAC. The conjugate antibody is common for all assays, and it targets neoepitopes generated after the formation of the MAC (alkaline phosphatase conjugated antihuman C5b-9) [13].

Another automated method for complement function analysis is the liposome enzymatic assay. Briefly, a liposome containing glucose-6-phosphate (G6P) and a hapten on its surface are used as the first reagent. A second IgM-containing reagent, which targets the hapten on the surface of the liposome, creates an artificial immune-complex. In the presence of patient sera, which are added after the formation of the aforementioned complex, the CP should be triggered when in contact with the immune-complex created by the reagents. The CP progresses up to the creation of the MAC, which is anchored on the surface of the liposome, creating holes in it. G6P is exposed after the liposome is lysed by the MAC, and an enzymatic reaction using nicotinamide adenine dinucleotide can be measured on several automated platforms, with absorbance measured at 340 nm (Fig. 23.4). This assay is FDA-approved for measurement of CP activity.

Complement component concentrations

Complement component concentrations are the second most common complement assessment after CH50 and AH50. Low complement can signal consumption and clearance of immune-complexes or inflammation, as seen in flares of autoimmune disease activity. C3 and C4 are routinely measured in most laboratories using nephelometric or turbidimetric approaches. Both nephelometric and turbidimetric measurements utilize the formation of antibody–antigen immune-complexes to determine protein concentrations. When the antigen to be measured, such as C1-INH, C3, C4, and C5, FB and FH are in equilibrium with the antibodies to the complement

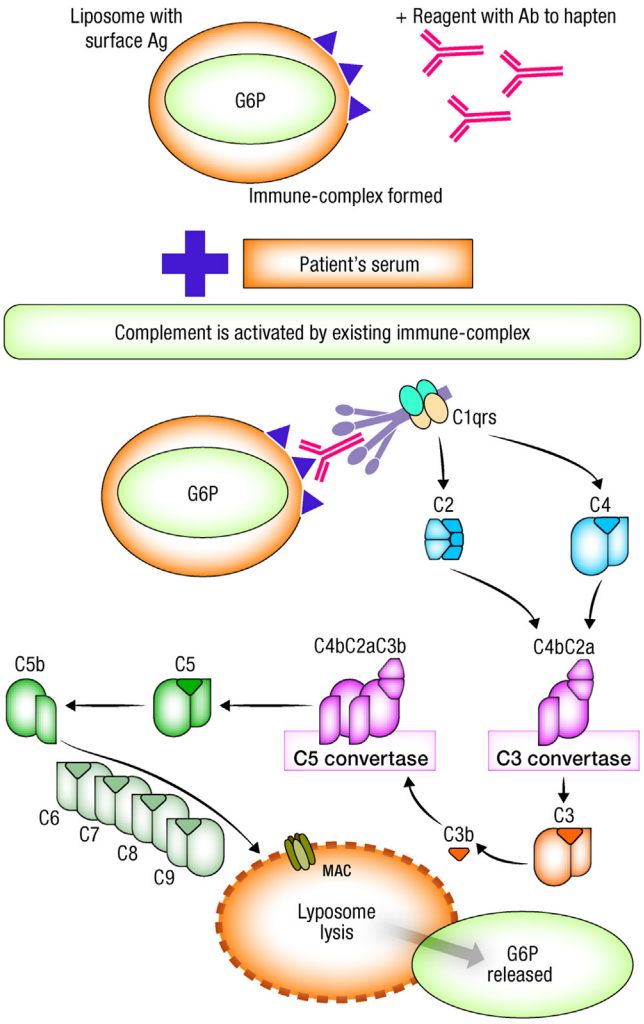


FIGURE 23.4 Illustration of complement functional assay to measure the activity of the classical pathway using a liposome colorimetric method. Reproduced with permission of the Mayo Clinic Foundation, retainer of the copyrights for publication.

component, immune-complexes are formed that can then be measured using light scattering approaches. Nephelometric methods quantitate an analyte based on the light scatter from immune-complexes at an angle, while turbidimetry allows for antigen quantitation based on changes in the transmission of light with an in-line detector. The outputs of both methods are directly proportional to concentrations of the complement component of interest.

It should be mentioned that particles, solvent, and macromolecules all scatter light and affect light transmission, which means lipoproteins and chylomicrons are problematic interfering substances for both nephelometry and turbidimetry. Lipemic samples can generate signal above the background despite lack of antigens of interest, and therefore cannot be tested. It is not common practice to ultracentrifuge specimens to remove lipemia because of the poor stability of the samples and potential risk of complement activation. Issues confounding the measurement of many complement components in serum are that the antibodies used by most diagnostic laboratories react with the parent, intact molecules, as well as many of their cleavage, activation fragments. This is common for C3, FB, and C4. C4 is also an acute phase reactant, and can double concentrations within 24 hours after trauma or acute inflammation. Therefore it should be kept in mind that the measurement returned by such methods will not distinguish if the component was cleaved in the patient, or not. For such questions, functional tests and analysis of the specific split products are indicated.

Manual methods for measuring complement component concentrations, such as RID, are still considered relevant, particularly for C2. RID measurement of antigen concentrations also relies on the formation of immune-complexes, as with the aforementioned methods; however, in this instance, a precipitin ring will be observed between the antiserum against a specific complement component mixed with the agarose gel, and the complement component in patient serum. The diameter between sample and precipitin ring will be compared with standards for quantitation. ELISA assays are not only set up as functional assays, but also commonly used for measurement of complement components, particularly for lower abundant proteins. In addition, ELISA approaches overcome some of the interferences observed with the turbidimetric and nephelometric approaches.

Complement activation products

Complement activation may be assessed by measurement of complement activation fragments or split products, which reflect the stimulation of any of the activating pathways. The main challenge for analysis of the activation fragments is to stabilize the sample, so that complement activation is inhibited in vitro. Plasma samples are required for the majority of the tests in this category, especially when performed by reference laboratories, because time to analysis and transportation play a significant role in accidentally activating the complement system in vitro.

The main methodologies available for analysis of activation fragments are ELISAs or fluid phase immunoassays using electrochemiluminescent detection approaches [14], which may use polyclonal or monoclonal antibodies against the many different unique epitopes to each of the specific fragments. Cross-reactivity is a challenge, and different assay manufacturers do not use antibodies for the same targets, making study results inconsistent and reference intervals very different, all of which should be taken into consideration when interpreting results.

Autoantibodies to complement components

Given the heterogeneity in autoimmune response to complement, the main methods used to detect autoantibodies to complement proteins are the aforementioned immunoassay strategies (e.g., ELISA and electrochemiluminescence). Analytes of interest stem from the identification of a role of anti-C1q autoantibodies in the pathogenesis of lupus nephritis (LN) and in hypocomplementemic urticarial vasculitis. Autoantibodies to C1q and C1-INH are usually measured by ELISA, and testing is most commonly performed in the setting of autoimmune diseases and acquired angioedema (AAE), respectively. Antibodies to FH are another common target for laboratory analysis. These autoantibodies have been measured with a standardized ELISA validated in an international multicenter study. When utilizing testing for complement autoantibodies, most assays are designed to detect only IgG autoantibodies. Of note, autoimmunity to complement components themselves can result in disease manifestations. This is particularly true of the autoantibodies associated with the kidney, such as C3 nephritic factors (C3Nefs) or autoantibodies to FH, which are also associated with atypical hemolytic syndrome and with AAE (antibodies to C1-INH).

C3Nefs are IgG and IgM autoantibodies that bind directly to C3 convertase of the AP. These molecules are very heterogeneous. In order to accurately measure C3Nefs, a stable C3 convertase must be generated in vitro; this would allow for autoantibody capture. This complex would not progress to cleave C3 and change into a C5 convertase. The gold standard laboratory method for C3Nefs testing is a hemolytic assay that measures whether patient-purified IgGs can stabilize preformed membrane-bound C3 convertase from the AP. Another approach is to measure the C3Nefs via surrogate, by indirect electrophoresis-based detection, which infers the presence of C3 convertase stabilizing molecules when C3 breakdown products are detected in serum. The latter, however, does not correlate with direct hemolytic assays according to proficiency testing reports. C4 nephritic factors are autoantibodies against the CP and LP C3 convertases, which require similar methodologies for measurement; however, these autoantibodies are very rare. These are currently manual, labor-intensive laboratory developed tests, only offered by very few reference laboratories.

Complement genetic testing

The complement serological tests are not always sensitive or specific enough to aid in diagnosis of the rare complement deficiencies or conditions with complement dysregulation such as atypical hemolytic uremic syndrome (aHUS) or PNH. Identifying if the defect is genetic or acquired is helpful in many conditions for treatment decisions. For instance, patients with a large deletion in FH are at higher risk for developing autoantibodies to FH. Patients without genetic variants have better prognostic outcomes in aHUS than the ones with complement variants, which are at higher risk for relapses. Genetic tests for complement-related genes are usually performed by next generation sequencing (NGS). Given that complement proteins have high homology, copy number variation and detection of large deletions is necessary in addition to NGS. Genetic analyses are beyond the scope of this chapter, but comprehensive information can be found in the references list [15–17].

Disorders associated with complement deficiency or dysregulation

Complement deficiencies

Increased susceptibility to infection was the first recognized clinical outcome of having a complement deficiency. Since those initial investigations, an increase in infection has been associated with deficiencies with almost every complement component. This ranges from the apparently more common and often benign deficiency in MBL to the severe outcomes of deficiencies in properdin. Individuals with complement component deficiencies, largely presenting in childhood, experience more frequent and severe infections involving pyogenic bacteria. Warning signs for a complement-mediated primary immunodeficiency include meningococcal meningitis in patients >5 years of age; other recurrent bacterial infections, especially *Pneumococcus*; autoimmune manifestations; angioedema without urticaria; and renal and ophthalmic inflammatory disorders. As a deficiency in any of a number of complement components that can result in similar infectious presentation, diagnosing the specific deficiency can be complicated.

The prevalence of MBL deficiency has been reported to be around 10% in European populations. Often without clinical symptoms, the most common presentations of disease are development of autoimmune disease or more frequent and severe infections. Deficiencies in C2 are most common in Caucasians, occurring at a rate of 1:10,000 to 1:20,000, resulting in autoimmune manifestations or infection, particularly streptococcal pneumonia. There are two forms of C2 deficiency: Type I C2 deficiency, found in over 90% of patients resulting from a complete lack of C2 protein synthesis (messenger RNA is present but no translation), and Type II C2 deficiency resulting from an intracellular protein secretion defect [18]. C4 deficiencies can also be associated with autoimmune or infectious presentations. C4 is encoded by two different genes in the

major histocompatibility complex on human chromosome 6. Seventy-five percentage of the population has two *C4A* and two *C4B* genes; however, the total sum of *C4A* and *C4B* genes in an individual can range from zero to eight or more copies, giving this protein a wide range of concentrations and an even wider range of functionality among the general population. Most of the partial C4 deficiencies are without consequence, although deficiency of *C4A* is associated with a 15% incidence of SLE.

Deficiencies in the AP or TP of complement are the most clinically significant for recurrent infections. Deficiencies in properdin, the only positive regulator in complement and the only component that is X-linked, result in severe and sometimes fatal infections. Similarly, deficiencies in components C3, C5, C6, C7, and C8 are strongly linked with susceptibility to pyogenic infections, particularly *Neisseria* [19]. Recommended testing for complement deficiencies is listed in Table 23.1.

Complement in autoimmune diseases

Disease presentations

Clearance of immune-complexes is an important housekeeping role of the complement system, mediated by activation of the CP and its PRM C1q. A close connection between complement deficiencies and autoimmune diseases exists, and complement is said to have a dual role. The first is that an impaired complement system is not as effective in clearing immune-complexes, and hence complement deficiencies are associated risk factors to development of autoimmune diseases, for example, SLE, primary Sjögren's syndrome (pSS), and RA [26]. The second role, in contrast to the first, is that the increased presence of large immune-complexes in autoimmune diseases demands large complement supply, and frequently depletes complement components for clearance of immune-complexes. Active states of disease are associated with hypocomplementemia, and increased ratio of activation fragments of the components of the CP. C1q deficiency is associated with 55%−75% of SLE cases. These patients present with several characteristics that may distinguish them from patients with sporadic SLE: early age of onset (~5 years of age), lower frequency of anti-dsDNA antibodies, prominent photosensitivity, and fewer renal symptoms [27,28]. In addition, the ratio of females to males is roughly equal for C1q deficiency-associated SLE, unlike sporadic SLE, which is more prevalent in females.

The renal manifestations of SLE, known as LN, are associated with more severe disease morbidity and mortality. In epidemiological studies, it was found that, among other factors, hypocomplementemia has prognostic significance for the development of LN. Autoantibodies against C1q have been reported in over 30% of SLE patients with LN. The presence of anti-C1q autoantibodies alone or in combination with anti-dsDNA and low complement has been shown to have strong association with renal involvement in SLE [29,30]. In addition to anti-C1q antibodies, autoantibodies to C3 and C3Nefs may have a role in LN [31]. CP activation is found in RA, polyarteritis nodosa, hypersensitivity vasculitis, rheumatoid vasculitis, SLE vasculitis, and mixed cryoglobulinemia, among many other conditions. In SLE, pSS, and RA, measurements of C3 and C4 have been used to monitor disease activity. A low result in C4 or both C3 and C4 is suggestive of active disease, and may help stratify patients' risks. Given that C3 is the most abundant complement component, and C4 is a protein coded by genes with different copy numbers, a decreased concentration below reference intervals is a significant finding. Although C3 and C4 measurements are widely available and have proved useful for most patients, a number of conditions can cause low C3 and C4, leading to lack of specificity.

Hereditary and acquired angioedema

Hereditary angioedema (HAE) is caused by a deficiency in the C1-INH, the PRM of the CP. C1-INH deficiency stems from heterozygous variants in the *SERPING1* gene (chromosome 11, q12-q13.1). Over 200 variants have been described, and the vast majority are inherited in an autosomal-dominant fashion. Links between specific variants and disease severity have not been identified; therefore the burden of disease may vary in individuals with similar pathogenic variants. HAE affects approximately 1 in 50,000−100,000 individuals. It is characterized by recurring episodes of unpredictable acute nonpitting edema attacks without urticaria or pruritus, predominantly affecting the cutaneous and/or mucosal tissues. The onset of disease symptoms usually occurs in the second decade of life. While the frequency of angioedema attacks is highly variable, most acute attacks occur at least once a month and last 1−5 days, incapacitating the patients for ~1−2 months per year.

The most debilitating manifestations of HAE include abdominal pain and vomiting due to gastrointestinal swelling or asphyxiation due to laryngeal swelling, but attacks can also occur in the face, hands, feet, or urogenital track. Attacks of the hands or feet are generally unilateral. Although acute attacks are spontaneous, numerous triggers have been reported; examples include stress, trauma, certain pharmacological exposures, infection, or inflammation. If HAE is unrecognized, abdominal attacks may lead to unnecessary surgical procedures. Furthermore, asphyxiation may be fatal and is the predominant cause of death in these patients.

TABLE 23.1 Laboratory testing indicated in several clinical conditions with complement involvement.

	Characteristics	First tier assays and findings	Second tier assays and findings	Genetic testing
Complement deficiencies	Individual components or control proteins are low or absent. Gene variants are usually involved.	CH50 and AH50 may be decreased.	If only CH50 is low: test for C1, C4, and C2 (function and concentrations). If only AH50 is low: test for FD, FB, and properdin. If both CH50 and AH50 are low: test for C3, C5, C6, C7, C8, and C9 (function and concentrations). If both CH50 and AH50 are normal: test for LP function.	Sequence gene coding for low serological results is recommended, but not mandatory, once the phenotypical presentation is characterized.
SLE, RA, and other autoimmune diseases. Monoclonal gammopathies and HCV infections	Increased immunoglobulin production (with or without cryoglobulinemia) or immune-complex formation will lead to CP over-activation. Deposits of immune-complexes in the kidneys are common.	CH50, C3, and C4. Common results in the setting of hypocomplementemia are low CH50 and low C4, with normal or low C3. Type II cryoglobulinemia presents with low C3 and very low or absent C4.	C1q function and concentration. Autoantibodies to C1q are associated with lupus nephritis. Activation products of C3 and C4 can provide estimates of disease activity in SLE.	Not usually required.
Immune-complex-mediated MPGN and C3 glomerulopathies	Rare kidney diseases with deposits on the glomerular basement membrane. Immune-complex-mediated MPGN results from infectious processes, autoimmune diseases, or monoclonal gammopathies, whereas C3 glomerulopathies are differentiated into C3 glomerulonephritis or dense deposit disease using electron microscopy.	CH50, AH50, C3, C4, factor B, factor H, and sC5b-9. Expected findings are decreased function and concentration, with increased activation fragments (sC5b-9).	Immune-complex-mediated MPGN: CP involvement may be evident when activation products (e.g., C4d) are tested. C4 nephritic factors have been reported. C3 glomerulopathies: strong involvement of AP. Activation fragments such as Bb and sC5b-9 as well as C3 nephritic factors and antibodies to FH are recommended tests.	Genetic causes of C3G include variants of C3 complement gene and regulatory proteins (*CFH*, *CFI*, *MCP*, *CFHR5*, and *CFHR3−1*) [20–25]. Allele variants for *CFH* and copy number variation for the *CFHR* cluster have been reported.
Age-related macular degeneration	Complement components C3, C5, and MAC have been found in eyes retina, contributing to oxidative stress and formation of drusen.	Aqueous fluid measurements usually include evaluation of cytokines (IL-6, IL-10, and VEGF), but there is no routine measurement of serum complement assays.	In the event of concomitant MPGN, testing for factor H and anti-FH autoantibodies may be of value.	Y402H polymorphism in *CFH* is a lead predictor of risk for AMD, followed by CFHR1−5, C3, CFI, CFB, C2, and C9.
Atypical hemolytic uremic syndrome	This heterogeneous, rare disorder with presentations of nonimmune hemolytic anemia, thrombocytopenia, and renal impairment has strong AP involvement.	A panel of assays including CH50, AH50, C3, C4, FB, Bb, FH, and sC5b-9 will provide information to discern between CP and AP involvement. Patients with aHUS had lower CH50, C3, and FB when compared with secondary non-aHUS TMA, and a	Other conditions such as TTP may activate the alternative pathway and result in increased levels of C3a, C5a, and sC5b-9, although not to the same extent as in aHUS. Normal FH, FI, CD46, FB, and C3 concentrations do not exclude aHUS diagnosis, since	Sequencing *CFH* and copy number variation for *CFHR1*, *CFHR3*, *CFHR4*, sequencing *CFB*, *CFI*, *C3*, *CD46*, and THBD.

(*Continued*)

TABLE 23.1 (Continued)

	Characteristics	First tier assays and findings	Second tier assays and findings	Genetic testing
		combination of assays using FB ≤ 20.9 mg/dL and CH50 ≤ 56% led to a sensitivity of 75% with increased specificity of 89%. Additional serological findings may include low AH50, low FH, or FI concentrations. Reduced C3 is found in only 30%–50% of patients with aHUS, TTP, and *Shiga toxin* HUS.	some variants may lead to a functionally impaired protein without a detectable change in concentration. In such instances, functional testing, when available, or genetic analysis can serve as an effective way to detect variants in complement pathway genes.	
Paroxysmal nocturnal hemoglobinuria	PNH is an acquired hematologic disorder characterized by nocturnal hemoglobinuria, chronic hemolytic anemia, thrombosis, pancytopenia, and, in some patients, acute or chronic myeloid malignancies. PNH affects erythroid, granulocytic, and megakaryocytic cell lines. The abnormal cells in PNH have been shown to lack (GPI)-linked proteins in erythroid, granulocytic, megakaryocytic, and, in some instances, lymphoid cells.	CH50, the Ham's test or flow-cytometry panel.	Flow cytometry can detect the presence or absence of GPI-linked proteins in granulocytes, monocytes, erythrocytes, and/or lymphocytes. A partial list of known GPI-linked proteins include CD14, CD16, CD24, CD55, CD56, CD58, CD59, C8-binding protein, alkaline phosphatase, acetylcholine esterase, and a variety of high-frequency human blood antigens.	Variants in the phosphatidylinositol glycan A gene have been identified consistently in patients with PNH, thus confirming the biological defect in this disorder.

Notes: Functional assays for factor B, factor D, and properdin are not available. MBL deficiency frequency is reported as 10% in individuals of European descent and frequently goes without clinical manifestations. Screening is indicated only in the context of symptoms, such as repetitive infections, when CH50 is normal. *AAE*, Acquired angioedema; *aHUS*, atypical hemolytic uremic syndrome; *AMD*, age-related macular degeneration; *AP*, alternative pathway; *CP*, classical pathway; *FB*, factor B; *FD*, factor D; *FH*, factor H; *FI*, factor I; *GPI*, glycophosphatidylinositol; *HAE*, hereditary angioedema; *LP*, lectin pathway; *MPGN*, membranoproliferative glomerulonephritis; *PNH*, paroxysmal nocturnal hemoglobinuria; *SLE*, systemic lupus erythematosus; *THBD*, thrombomodulin; *TMA*, thrombotic microangiopathy; *TTP*, thrombotic thrombocytopenic purpura; *VEGF*, vascular endothelium growth factor.

TABLE 23.2 Laboratory features consistent with hereditary and acquired angioedema subtypes.

	Type I HAE	Type II HAE	AAE
C1-INH concentration	Low	Normal/high	Low
C1-INH function	Low	Low	Low/normal
C4 concentration	Low	Low	Low
C1q concentration	Normal	Normal	Low
Anti-C1-INH antibodies	Absent	Absent	Present
C3 concentration	Normal	Normal	Normal
Family history	Yes	Yes	No

AAE, Acquired angioedema; *C1-INH*, C1 esterase inhibitor; *HAE*, hereditary angioedema.

It has been demonstrated that HAE attacks may be attributed to the role of C1-INH in the contact activation pathway, where it controls bradykinin release through inhibition of factor XIIa [32]. In HAE, there is insufficient C1-INH to negatively regulate bradykinin release and stop angioedema attacks from occurring. The mechanism of HAE attacks is distinct from an allergic angioedema, as it is not mediated by histamine release via mast cell activation. Therefore HAE patients are unresponsive to antihistamines or corticosteroids.

There are two main types of HAE that are attributed to C1-INH deficiency (Type I HAE) or dysfunction (Type II HAE); impacted individuals can have C1-INH activities ranging from $<20\%$ to 50% of normal. Type I HAEs, representing approximately 85% of patients, are associated with low circulating concentrations of C1-INH, leading to a concomitant decrease in C1-INH function. In Type II HAEs, normal or elevated concentrations of functionally inactive C1-INH are produced. The relative proportion of Type I and Type II HAEs may differ based on geographical location. A third HAE subtype with unknown prevalence termed "HAE with normal C-INH" has been described. Although poorly characterized, a minority of these patients is known to harbor a mutation in factor XII; the disease origin for the remainder of patients remains unknown. Factor XII is the zymogen form of factor XIIa, and plays a key role in bradykinin production as part of the contact system.

Angioedema due to C1-INH deficiency can also be acquired during adulthood in the fifth decade of life or later [33]. The prevalence of AAE is extremely low and is estimated at 10% of HAE [34]. AAE is frequently associated with monoclonal gammopathies or lymphoproliferative disease, different types of cancer, and autoimmune diseases. AAE may be caused by development of anti-C1-INH autoantibodies that act to reduce the functional activity or increase the catabolism of C1-INH.

For patients exhibiting symptoms associated with HAE, evaluation of pertinent family history in combination with laboratory results for C1-INH function and concentration, C4 concentration, and C1q concentration can assist in HAE diagnosis and determination of HAE type (Table 23.2). Identification of low C1-INH function and low C4 concentration supports the diagnosis of HAE, and was found to have 98% specificity toward C1-INH deficiency and a negative predictive value of 95% [35]. C4 is decreased owing to excessive consumption through CP in the absence of inhibition by C1-INH. Due to low disease prevalence, false positive results are common; therefore repeated testing is recommended to confirm findings [36]. The C1-INH concentration assay can be used to distinguish Type I HAE, with low C1-INH concentration, from Type II HAE characterized by normal or elevated concentration. Furthermore, serum C1q concentrations can be used to differentiate HAE from AAE forms of angioedema, as the latter is characterized by decreased C1q antigen concentration [37] and autoantibodies against C1-INH. Genetic analysis for *SERPING1* variants status may also help exclude HAE [33].

Atypical hemolytic uremic syndrome

The HUS is a form of thrombotic microangiopathy (TMA) characterized by a triad of nonimmune hemolytic anemia, thrombocytopenia, and renal impairment. This TMA is often due to the infectious *Shiga* toxin producing *Escherichia coli* with the most common serotype 0157: H7, a clinical presentation that includes watery or bloody diarrhea. A more rare form of HUS has become more prevalent in recent years in individuals without known *Shiga toxin*-producing infections. These individuals usually had recurrent disease associated with a more severe clinical course and worse outcomes. Patients fitting this clinical phenotype are diagnosed as having atypical HUS

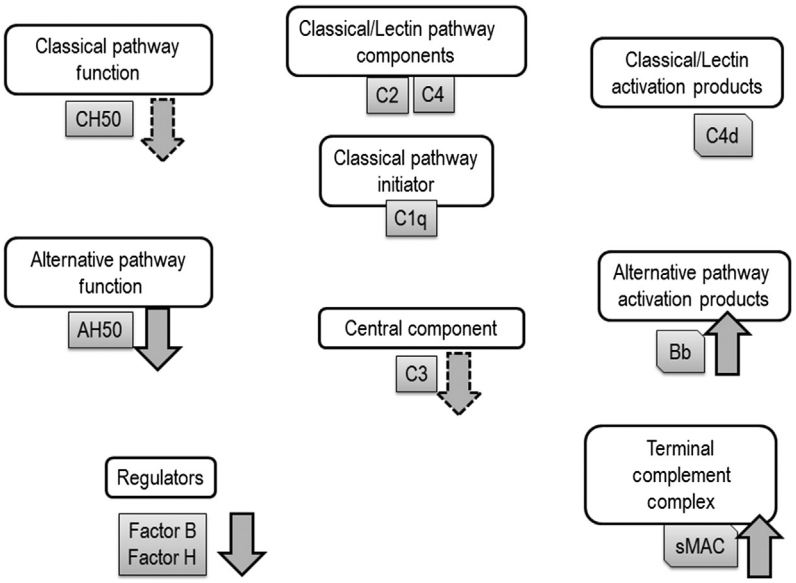


FIGURE 23.5 Expected pattern of findings for alternative pathway dysregulation. Complement alternative pathway function will be decreased, as well as concentrations of the components of the alternative pathway, due to increased in vivo consumption of those early components or regulators, and there is accumulation of the activation fragments, suggesting in vivo cleavage above reference intervals. C3 may or may not be decreased. If it is decreased, it may also impact the CH50 assay, which relies on the C3 available to measure the classical pathway function. A combination of findings or only a few of the analytes may be decreased/increased, as the sensitivity/specificity of the assays may vary with methodology used, and time of collection in relation to symptoms presentation. Evaluating for the presence of autoantibodies and genetic variants is also recommended.

or complement-mediated TMA. The incidence of the far less common atypical form of HUS (aHUS) in Europe is estimated at 2 per million population per year. aHUS is a rare and highly heterogeneous disorder that may appear during childhood or adulthood. The majority of patients incur irreversible renal damage, progress toward end-stage renal disease, or die within the first year of diagnosis when treated with plasma exchange or infusion. Over 80% of aHUS cases are not linked to family history of the disease, and therefore are considered sporadic. Approximately, half of sporadic cases are known to occur subsequent to a trigger event such as common infections or pregnancy. The inherited form of the disease is associated with a particularly poor outcome and has complex underlying genetics with frequent and rare variants and copy number variations of complement genes [38].

aHUS has also been called complement-mediated HUS and is a result of hyperactivation of the AP. The pathophysiologic mechanism involves increased continuous spontaneous hydrolysis of C3 to C3b, leading to tissue deposition of C3b, MAC formation, and subsequent tissue injury. This may be a result of genetic variants in complement-related genes or acquired autoantibodies to complement components. Over 100 variants in complement pathway proteins have been described, which account for $\sim$50%–60% of aHUS cases; however, additional causative mechanisms are yet to be uncovered. The known variants include gain-of-function variants of complement pathway activators (complement C3 and FB) and loss-of-function variants of regulators (FH, CD46, FI, and thrombomodulin). Pathogenic variants are found in only about half of the cases of aHUS. The penetrance of most variants is low, suggesting the requirement for other genetic or acquired factors for phenotypic expression of the disease, for instance, the association of FH autoantibodies with a deletion of *CFHR1* and *CFHR4* in addition to variable copies of deleted *CFHR1* and *CFHR3* genes. The autoantibodies inhibit FH function by binding and blocking the C-terminus, thereby impairing its ability to bind endothelial cell surfaces, sialic acids, and C3b.

In patients with clinical presentation characteristic of aHUS, lab features can be used for diagnosis and as a guide for treatment decisions [39]. In an initial workup, a distinction must be made between aHUS and infection-induced HUS or thrombotic thrombocytopenic purpura (TTP), as the disease management differs based on these distinctions, in particular with regard to the use of eculizumab, approved for aHUS treatment but not TTP. Immunodetection of the *Shiga toxin* in stool is consistent with infection-induced HUS. A laboratory finding of a significant decrease in ADAMTS13 activity ($<$10% of normal plasma) is consistent with TTP and can be used to distinguish it from aHUS. The serum complement profile for aHUS should reflect preferential activation of the AP and complement consumption with accumulation of the activation fragments from the AP, without CP/LP involvement (Fig. 23.5). In a recent study, patients with aHUS had lower CH50, C3, and FB, when compared with secondary non-aHUS TMA, and a combination of assays using FB $\leq$20.9 mg/dL and CH50 $\leq$56% led to a sensitivity of 75% with increased specificity of 88.9% and a diagnostic odds ratio of 24 [40]. Additional serological findings may include low AP function (AH50) or low FH or FI concentrations [41]. Reduced C3 is found in only 30%–50% of patients with aHUS, TTP, and *Shiga toxin*-related HUS. Other conditions such as TTP may activate the AP as well, and result in increased levels of C3a, C5a, and sC5b-9, although not to the same extent as in aHUS.

TABLE 23.3 Serologic assays that may be used as a panel to evaluate AP dysregulation in aHUS.

Test	Utility
CH50	Measurement of total complement function, triggered by an immune-complex. The assay requires complete activation of the classical pathway all the way to the formation of the membrane attack complex. If any of the components are decreased, CH50 measurement can be affected.
C4	C4 is a component of the classical and lectin pathways exclusively. If C4 concentrations are reduced, it suggests those pathways are involved. It does not participate in the alternative pathway. Standard assays for C4 do not distinguish between the parent component and its activation products. C4 is also an acute phase reactant, with concentrations that can double within 24 hours after trauma or acute inflammation. The measurement returned by such methods will not distinguish if the component was cleaved in the patient, or if a decreased result is due to a C4 deficiency.
C4d	C4d is a stable activation product of C4, further indicating classical or lectin pathway dysregulation, and C4 consumption.
AH50	Measurement of alternative pathway function. The most common method available utilizes an ELISA plate coated with lipopolysaccharides and proprietary buffers to block activation of the classical and lectin pathways. The complement proteins in patient's serum are activated by the presence of lipopolysaccharides. C3 hydrolysis initiates the cascade that ends with the formation of the sC5b-9 terminal complement complex. The detection antibody used for this test targets a neoepitope, only exposed on the sC5b-9 complex.
C3	The central component of the classical, lectin, and alternative pathways. This assay and other standard assays for C3 do not distinguish between the parent component and its activation products. Since C3 is the most abundant complement protein, a decreased result is a significant finding.
FB	Activator of the alternative pathway, binding to deposited C3b or hydrolyzed C3. Requires activation by factor D to form the alternative pathway C3 convertase complex, which cleaves additional C3 to C3b.
Bb	FB is activated by factor D and generates two activation fragments, Ba and Bb. Bb will bind to a deposited C3b molecule and create the AP C3 convertase, C3bBb.Measurement of soluble Bb indicates activation of the alternative pathway by FB and factor D.
FH	Complement FH regulates the alternative pathway of complement activation to prevent damage to self. It binds to glycosaminoglycans on cell surfaces through its central part (short consensus repeats SCRs 6–8) and its carboxyl-terminus (SCRs 19–20). The amino-terminal part competes with FB for binding to C3b and acts as a cofactor for factor I to inactivate C3b to iC3b, whereas the C-terminus binds to C3b and its cleavage fragment C3d. FH also enhances the dissociation of the C3bBb complex, limiting the alternative pathway activity.
sMAC	The terminal complement complex or soluble membrane attack complex is formed by C5b, C6, C7, C8, and C9 (C5b-9). The term "soluble" refers to the MAC in circulation in the fluid phase, when bound to one of its control proteins, vitronectin or clusterin. The MAC is responsible for anchoring on the cell membrane and promoting cell lysis when the control proteins are released from its structure. The ELISA assay available utilizes a capture monoclonal antibody to the C9 ring of sMAC, and detection antibodies conjugated with HRP that bind to antigens of the sMAC complex. High concentrations of sMAC are associated with dysregulation of alternative pathway.
Antibodies to FH	The autoantibodies inhibit FH function by binding and blocking the C-terminus of the molecule, thereby impairing its ability to bind endothelial cell surfaces, sialic acids, and C3b.

aHUS, Atypical hemolytic uremic syndrome; *AP*, alternative pathway; *FB*, factor B; *FH*, factor H; *HRP*, horseradish peroxidase; *MAC*, membrane attack complex; *sMAC*, soluble membrane attack complex.

Normal FH, FI, CD46, FB, and C3 concentrations do not exclude aHUS diagnosis, since some variants may lead to a functionally impaired protein without a detectable change in concentration. In such instances, functional testing, when available, or genetic analysis can serve as an effective way to detect variants in complement pathway genes. Several groups have performed retrospective analysis and demonstrated panels with high specificity for complement-mediated TMA [42,43], aHUS, and C3 glomerulonephritis (C3GN). An example of analytes to be used in such panels is described in Table 23.3.

Paroxysmal nocturnal hemoglobinuria

PNH is a rare clonal bone marrow disorder occurring at a population-level incidence of 5–10 in a million and affecting primarily young adults [44]. It is enriched in certain populations, including patients with intravascular hemolysis or hemoglobinuria, bone marrow failure, or

unexplained thrombosis, which are diagnostic features for PNH. PNH develops as a result of loss of GPI-anchors from cell membranes, resulting in untethering of membrane inhibitors of complement, thereby rendering the cells more susceptible to complement-mediated destruction.

RBCs express the essential complement regulatory protein CD59 on the plasma membrane via a GPI anchor, and its presence inhibits the formation of the MAC, thereby protecting cells from autologous complement-mediated destruction. In PNH, the synthesis of phosphatidylinositol tails required for anchorage of CD59 to the plasma membrane of RBCs is impaired due to a somatic mutation in the X-linked gene, phosphatidylinositol glycan class A. The absence of the critical CD59 leads to abrogation of inhibitory mechanisms, allowing complement activation, transversion of the MAC through the lipid bilayer, and subsequent destruction of RBCs. Consequently, PNH patients experience chronic intravascular hemolysis and numerous associated morbidities including hemolytic anemia, venous/arterial thrombosis, and smooth muscle dystonia. In addition, bone marrow failure can progress to aplastic anemia in the most severe cases. Occurring in approximately half of patients, thromboembolism is the most severe complication of PNH, and the primary cause of death. Prior to the approval of eculizumab, supportive therapies such as blood transfusions and stem cell transplantation were the mainstays of treatment in these patients, and were associated with poor quality of life and significant morbidities. With the emergence of novel therapies, the outlook for PNH is much improved.

Flow cytometry-based quantification of GPI-linked proteins in granulocytes, monocytes, RBCs, and/or lymphocytes is used as a diagnostic assay for PNH, an improvement which avoiding the problems associated with RBC-based diagnostic methods (Ham's test), in which recent hemolytic episodes or recent transfusions can give false-negative results. Flow cytometry-based methods are strongly preferred over the alternative methods, as flow provides superior specificity, quantitation, and sensitivity. A partial list of known GPI-linked proteins includes CD14, CD16, CD24, CD55, CD56, CD58, CD59, C8-binding protein, alkaline phosphatase, acetylcholine esterase, and a variety of high-frequency human blood antigens.

Complement in C3 glomerulopathies

C3 glomerulopathies is an umbrella term to define C3 deposition in the kidney without the presence of immunoglobulins, as it is observed in C3GN and dense deposit disease (DDD). This is in contrast to immune-complex-mediated membranoproliferative glomerulonephritis (MPGN). The C3G conditions are thought to be a result of complement dysfunction in the AP and differ from classical postinfectious glomerulonephritis, where immune-complexes are present and CPs or LPs are involved. The differentiation between C3GN and DDD is made using electron microscopy, according to the pattern observed in the glomerular basement membrane. Despite phenotypic differences, these glomerular diseases share dysfunction of the AP as the defining pathophysiology. As such, the diagnostic approach to these disorders should be similar. The causal abnormality in the complement pathway, genetic or acquired, may not be clear until an abrupt event, such as infection or pregnancy, activates complement exponentially, in an uncontrolled manner.

Genetic causes of C3G include variants of C3 and regulatory proteins (*CFH, CFI, MCP, CFHR5*, and *CFHR3-1*). Acquired forms of the disease are associated with the presence of C3Nefs that bind and stabilize the C3 convertase, increasing its half-life almost 10-fold, from seconds to minutes or hours, and blocking its dissociation; therefore the enzyme is active continuously, promoting complement activation. This will in turn lead to C3 consumption and decreased serum C3 concentrations. There have also been reports of C4 nephritic factors that act through a similar mechanism, stabilizing the CP and LP C3 convertases [45].

In the context of renal disease, reduced C3 is found in 40%−75% of patients with C3G and approximately 60% of atypical postinfectious glomerulonephritis. Significantly decreased C3 concentrations may be found in deficiencies of FH, FI, or the presence of C3Nefs. Reduced concentrations of FB are associated with increased concentrations of Bb and Ba activation fragments, and are consistent with AP C3 consumption. AH50 is commonly decreased. Anti-FH antibodies should also be assessed, and the sC5b-9 is elevated in 9% of DDD patients and 50% of C3GN patients. Conversely, a normal FB concentration and reduced C4 may reflect CP activation if genetic deficiency of C4 is excluded, and point the diagnosis toward immune-complex-mediated MPGN.

Complement and other conditions

Monoclonal gammopathies

Autoimmune diseases often result in polyclonal elevation of the immunoglobulins repertoire, and the relationship of the excess of immune-complexes with the complement system has been reviewed. Monoclonal gammopathies may have a similar impact on the activation of the complement system, since the monoclonal proteins may also deposit in organs and tissues (e.g., light-chain amyloidosis, multiple myeloma, and kidney). Tissue biopsy

immunohistochemistry studies often reveal the presence of immunoglobulin deposits as well as complement C3b and C4. The monoclonal immune-complexes may trigger activation of the CP and the TP. Patients with renal damage owing to the accumulation of the monoclonal protein in the glomeruli are classified with monoclonal gammopathies of renal significance, a reclassification compiling all renal disorders caused by a monoclonal immunoglobulin.

Cryoglobulinemia

Cryoglobulins are proteins that precipitate when cooled and dissolve when heated. In most cases, monoclonal cryoglobulins are IgM or IgG. They are classified as follows: type I (monoclonal: IgG, IgM, IgA, or rarely monoclonal light chains); type II (mixed: two or more immunoglobulins, of which one is monoclonal); and type III (polyclonal: in which no monoclonal protein is found). The monoclonal protein or mixed immunoglobulins precipitate in cooler vessels, such as earlobes and extremities, causing damage due to poor perfusion leading to infarction or ischemia, leg cramps, acrocyanosis, Raynaud's phenomenon, ulcers, and purpura. The temperature at which the cryoglobulin precipitates is more important than the amount of protein. Most commonly, mixed cryoglobulins (type II) consist of monoclonal IgM with rheumatoid factor activity and polyclonal IgG. Type II cryoglobulinemia is notably likely to present with complement abnormalities, with low concentrations of C3 and very low or absent C4. The monoclonal proteins deposition in tissues or the precipitation of cryoglobulins may activate the complement system via CP, LP, or AP, leading to low results being reported for CH50, AH50, or LP functional assays. Patients with mixed cryoglobulinemia frequently have vasculitis, glomerulonephritis, or lymphoproliferative or chronic infectious processes, like hepatitis B or hepatitis C infections. Measurement of low C3 and low/absent C4 are characteristic in type II cryoglobulinemia.

Hepatitis C virus infections

Viruses have been shown to attenuate complement activation, and interactions of hepatitis C virus (HCV) and the host immune surveillance system may play a role in viral persistence. Serum C3 is depleted in HCV-infected cirrhotic patients, and complement activation has also been reported, with decreased C4 activity in chronically infected HCV patients. In addition, HCV infection is associated with mixed cryoglobulinemia (type II cryoglobulinemia), and the HCV core antigen has been detected in cryoprecipitates. There is a structural similarity between C1q and HCV, which could explain the finding. In addition to CP involvement, LP is involved as MBL binds to HCV envelope proteins and also contributes to complement activation. The HCV infection with cryoglobulinemia is complicated in the presence of B-cell proliferation and extrahepatic disease manifestations such as leukocytoclastic vasculitis, arthropathy, neuropathy, and MPGN [46,47]. In HCV-infected individuals, testing for the presence of cryoglobulins is recommended and may suggest development of extrahepatic complications [48].

Complement in age-related macular degeneration

Age-related macular degeneration (AMD) is the leading cause of irreversible blindness in the developed world [49] A variant in *CFH*, Y402H, was found to be the lead predictors of risk for AMD [50–52]. Since that initial connection, polymorphisms in a number of other complement genes have been connected to AMD [53,54]. In addition to *CFH*, polymorphisms in *CFHR1–5*, *C3*, and rare variants of *CFI* [55,56] have been described to be risk factors in AMD, whereas variants in *CFB*, *C2*, and *C9* were found to be protective, and associated with decreased risk for AMD [54].

Complement components C3, C5, and MAC were actually found in the eye before the genetic connection was made [57]. The finding of the complement components, as well as other inflammatory markers, in the drusen lesions of AMD have reinforced the notion that AMD is driven by an inflammatory response with improperly controlled complement activation. The area of the eye where the drusen and complement deposition is found is, by definition, an area of possible high oxidative stress, which is believed to be a key trigger for the complement activations. The current treatments for AMD center around the inhibition of the provascular signaling molecule vascular are endothelium growth factor [58], but recent work has shown that the efficacy of the treatment is modulated by complement [59]. Many of the treatments on the horizon for AMD now focus on modulation of the AP of complement [60,61]. Currently, complement testing is not common for AMD. Testing for the genetic polymorphisms associated with AMD has been suggested as a diagnostic predictor of AMD risk, but without clear treatment stratification, the utility of such testing has been questioned. With the interest in complement-targeting therapeutic in AMD, the likelihood that complement analysis will be utilized in the future is high.

Complement therapeutics

Therapeutics for hereditary angioedema

The primary treatment goals for Type I and II HAEs are to reduce disease morbidity through reduction of the frequency, duration, and severity of angioedema attacks, and to avoid mortality by preventing occurrence or

progression of laryngeal edema. There are currently five commercially available, FDA-approved therapies that effectively treat acute HAE attacks, including C1-INH replacement therapies and drugs targeting the bradykinin pathway. C1-INH replacement therapy restores functional C1-INH concentrations to the normal range and can be derived in two ways: through purification from plasma (Berinert, CSL Behring GmbH, and Cinryze, Shire ViroPharma Inc.) or production of recombinant protein (Ruconest, Salix Pharmaceuticals). Bradykinin pathway-targeting drugs include the plasma kallikrein inhibitor, ecallantide (Kalbitor, Shire Pharmaceutical), and the bradykinin B2-receptor antagonist, icatibant (Firazyr, Shire Pharmaceutical). Ecallantide prevents kallikrein-mediated bradykinin release from high molecular weight kininogen, thereby reducing swelling. Icatibant, acting downstream of ecallantide, renders the bradykinin B2 receptor inaccessible to bradykinin. Of note, Berinert, Cinryze, and Firazyr may be self-administered by the patients or by their family members, thereby reducing the burden of disease by allowing for an earlier intervention and preventing unnecessary healthcare expenditures by averting emergency room visits and hospitalizations.

Although there are significant differences in treatment approaches for AAE versus HAE, there is evidence that AAE patients may also benefit from C1-INH replacement therapy. While effective in some cases, the rapid turnover rate of C1-INH in some AAE patients may require higher doses, or they may be resistant to treatment with plasma-derived C1-INH altogether. Efficacy of icatibant, and on a smaller scale, of ecallantide has also been demonstrated in AAE.

The C5 inhibitor

Eculizumab (Soliris, Alexion Pharmaceuticals), a humanized monoclonal hybrid IgG2/4κ antibody therapeutic directed against C5, is FDA-approved for treatment of PNH and aHUS, and has orphan-drug designation status for treatment of refractory myasthenia gravis and neuromyelitis optica. Eculizumab inhibits the terminal complement pathway through simultaneous blockade of the generation of the potent prothrombotic and pro-inflammatory molecule, C5a, and the formation of the MAC initiator, C5b [62,63]. Since all paths of the complement cascade converge at the point of C5 activation, eculizumab is thought to have a broad potential application and is being clinically evaluated in many other disorders with complement overactivation [64].

In PNH, eculizumab has become the standard of care, proving to be a safe and effective therapy with long-lasting effects, potentially enabling patients to become transfusion-independent and extending their survival. Eculizumab dosing regimen prescribed for an average adult diagnosed with PNH is 600 mg weekly for the first 4 weeks, followed by 900 mg for the fifth dose 1 week later; then, 900 mg every 2 weeks thereafter. In aHUS, it is prescribed for an average adult at 900 mg weekly for the first 4 weeks, followed by 1200 mg for the fifth dose 1 week later, then 1200 mg every 2 weeks thereafter. As compared with plasma exchange, eculizumab may facilitate better disease control and tolerability, and can be beneficial as a first-line treatment in children who have high risk of complications from plasma exchange and patients with recurrent aHUS (with or without transplantation).

The drawbacks of eculizumab therapy are associated with its potentially life-threatening side effects, variations in response profiles, and the cost of treatment. Patients treated with eculizumab are at an increased risk of susceptibility toward life-threatening infections such as *Neisseria meningitides*; to prevent these, vaccinations and, in some cases, prophylactic antibiotic treatment are recommended. In addition, PNH patients harboring a rare polymorphism of the complement receptor 1 gene were found to be more likely to remain transfusion dependent in spite of eculizumab therapy, and exhibited suboptimal response to eculizumab [65]. As another major drawback, being potentially a lifelong therapy with a price tag of nearly $6000 per 300 mg vial, the cost of eculizumab may limit its use in routine clinical practice worldwide and may amount to ~$100,000 or ~$500,000 annually in children or adults, respectively [66].

Monitoring eculizumab therapy can be accomplished by clinical monitoring of the patients undergoing therapy. In PNH, clinical parameters used to assess treatment outcome include stable hemoglobin concentrations with reduction of hemolysis, decreased blood transfusion requirements, improved quality of life, and reduced frequency of thromboembolic events. In aHUS, the use of eculizumab is usually initiated after poor response to plasma exchange. Successful response to eculizumab therapy includes normalization of thrombocytopenia, resolution of the TMA, and improvement of renal function measured by creatinine or estimated glomerular filtration rate. However, if the clinical response to therapy is only partial, for example, with normalization of some of the parameters but not others, one may wonder if eculizumab is adequately blocking C5. In those situations, the availability of laboratory testing can be helpful to guide whether a dose increase or shorter infusion intervals of eculizumab should be considered.

On the other end of the spectrum, it has been proposed that a group of patients without pathogenic variants and normalization of all hematologic and renal parameters (no evidence of microangiopathic hemolytic anemia, no thrombocytopenia, and stable renal function for over 3 months), could benefit from dose deescalation or therapy discontinuation. For these patients, assessment of

complement blockage status in conjunction with disease monitoring would be valuable. Complement activity of <10% for the AH50 assay or results below the reference intervals for C5 function or CH50 on the liposome methods are indicative of therapeutic concentrations of eculizumab >100 mcg/mL, and seem like suitable assays to monitor eculizumab therapy. C5 concentrations may be measured along with any of the functional tests, especially in cases of poor response to eculizumab, and it will help exclude a complement C5 deficiency; in the presence of eculizumab, C5 concentrations will be elevated [67]. Importantly, nephelometric or turbidimetric assays to measure C5 concentration use antibodies as reagents that may not distinguish between free and eculizumab-bound C5; therefore C5 concentration assays alone should be used to monitor complement control. Eculizumab quantitation may also be measured using immunoassay or mass spectrometric techniques for therapeutic drug monitoring.

Future directions

The indications for complement testing have expanded beyond the traditional evaluation of complement deficiencies to a series of other indications that include kidney diseases, TMAs, and autoimmune diseases. With complement therapies gaining the spotlight, the clinical laboratory specialized in this testing has regained relevance and may be able to contribute significantly to patient care with a diverse array of tests, targeting the complement system.

Acknowledgments

The author would like to thank colleagues Ashley Frazer-Abel, Ph. D. (University of Colorado Denver, CO), Lusia Sepiashvili, Ph.D. (The Hospital for Sick Children, Toronto, CA), and Michael Mbughuni, Ph.D. (US Department of Veterans Affairs Hospital, Minneapolis, MN), for their contributions to the study of the complement system and for writing a thorough review with the author in 2016 that served as the basis for this chapter.

References

[1] G. Bajic, S.E. Degn, S. Thiel, G.R. Andersen, Complement activation, regulation, and molecular basis for complement-related diseases, EMBO J. 34 (2015) 2735–2757.

[2] N.S. Merle, S.E. Church, V. Fremeaux-Bacchi, L.T. Roumenina, Complement system Part I - molecular mechanisms of activation and regulation, Front. Immunol. 6 (2015) 262.

[3] N.S. Merle, R. Noe, L. Halbwachs-Mecarelli, V. Fremeaux-Bacchi, L.T. Roumenina, Complement system Part II: role in immunity, Front. Immunol. 6 (2015) 257.

[4] M. Kolev, G. Le Friec, C. Kemper, Complement--tapping into new sites and effector systems, Nat. Rev. Immunol. 14 (2014) 811–820.

[5] K. Li, S.H. Sacks, W. Zhou, The relative importance of local and systemic complement production in ischaemia, transplantation and other pathologies, Mol. Immunol. 44 (2007) 3866–3874.

[6] A.J. Szalai, A. Agrawal, T.J. Greenhough, J.E. Volanakis, C-reactive protein: structural biology and host defense function, Clin. Chem. Lab. Med. 37 (1999) 265–270.

[7] R. Ghai, P. Waters, L.T. Roumenina, M. Gadjeva, M.S. Kojouharova, K.B. Reid, et al., C1q and its growing family, Immunobiology. 212 (2007) 253–266.

[8] Y.H. Kang, L.A. Tan, M.V. Carroll, M.E. Gentle, R.B. Sim, Target pattern recognition by complement proteins of the classical and alternative pathways, Adv. Exp. Med. Biol. 653 (2009) 117–128.

[9] A. Frazer-Abel, L. Sepiashvili, M.M. Mbughuni, M.A. Willrich, Overview of laboratory testing and clinical presentations of complement deficiencies and dysregulation, Adv. Clin. Chem. 77 (2016) 1–75.

[10] H.D. Manthey, T.M. Woodruff, S.M. Taylor, P.N. Monk, Complement component 5a (c5a), Int. J. Biochem. Cell Biol. 41 (2009) 2114–2117.

[11] P.N. Nesargikar, B. Spiller, R. Chavez, The complement system: history, pathways, cascade and inhibitors, Eur. J. Microbiol. Immunol. (Bp). 2 (2012) 103–111.

[12] B. Nilsson, K.N. Ekdahl, Complement diagnostics: concepts, indications, and practical guidelines, Clin. Dev. Immunol. 2012 (2012) 962702.

[13] T.E. Mollnes, T.S. Jokiranta, L. Truedsson, B. Nilsson, S. Rodriguez de Cordoba, M. Kirschfink, Complement analysis in the 21st century, Mol. Immunol. 44 (2007) 3838–3849.

[14] A.R. Morgan, C. O'Hagan, S. Touchard, S. Lovestone, B.P. Morgan, Effects of freezer storage time on levels of complement biomarkers, BMC Res. Notes 10 (2017) 559.

[15] A. Servais, L.H. Noel, L.T. Roumenina, M. Le Quintrec, S. Ngo, M.A. Dragon-Durey, et al., Acquired and genetic complement abnormalities play a critical role in dense deposit disease and other C3 glomerulopathies, Kidney Int. 82 (2012) 454–464.

[16] V. Fremeaux-Bacchi, F. Fakhouri, A. Garnier, F. Bienaime, M.A. Dragon-Durey, S. Ngo, et al., Genetics and outcome of atypical hemolytic uremic syndrome: a nationwide french series comparing children and adults, Clin. J. Am. Soc. Nephrol. 8 (2013) 554–562.

[17] F. Bu, T. Maga, N.C. Meyer, K. Wang, C.P. Thomas, C.M. Nester, Comprehensive genetic analysis of complement and coagulation genes in atypical hemolytic uremic syndrome, J. Am. Soc. Nephrol. 25 (2014) 55–64.

[18] C.A. Johnson, P. Densen, R.A. Wetsel, F.S. Cole, N.E. Goeken, H.R. Colten, Molecular heterogeneity of C2 deficiency, N. Engl. J. Med. 326 (1992) 871–874.

[19] A.S. Grumach, M. Kirschfink, Are complement deficiencies really rare? Overview on prevalence, clinical importance and modern diagnostic approach, Mol. Immunol. 61 (2014) 110–117.

[20] M.A. Abrera-Abeleda, C. Nishimura, K. Frees, M. Jones, T. Maga, L.M. Katz, et al., Allelic variants of complement genes associated with dense deposit disease, J. Am. Soc. Nephrol. 22 (2011) 1551–1559.

[21] M.A. Abrera-Abeleda, C. Nishimura, J.L. Smith, S. Sethi, J.L. McRae, B.F. Murphy, et al., Variations in the complement regulatory genes factor H (CFH) and factor H related 5 (CFHR5) are

associated with membranoproliferative glomerulonephritis type II (dense deposit disease), J. Med. Genet. 43 (2006) 582−589.
[22] F.C. Fervenza, R.J. Smith, S. Sethi, Association of a novel complement factor H mutation with severe crescentic and necrotizing glomerulonephritis, Am. J. Kidney Dis. 60 (2012) 126−132.
[23] D.P. Gale, E.G. de Jorge, H.T. Cook, R. Martinez-Barricarte, A. Hadjisavvas, A.G. McLean, et al., Identification of a mutation in complement factor H-related protein 5 in patients of cypriot origin with glomerulonephritis, Lancet. 376 (2010) 794−801.
[24] C. Licht, S. Heinen, M. Jozsi, I. Loschmann, R.E. Saunders, S.J. Perkins, et al., Deletion of Lys224 in regulatory domain 4 of factor H reveals a novel pathomechanism for dense deposit disease (MPGN II), Kidney Int. 70 (2006) 42−50.
[25] S. Sethi, R.J. Smith, J.J. Dillon, F.C. Fervenza, C3 glomerulonephritis associated with complement factor B mutation, Am. J. Kidney Dis. (2014).
[26] K.N. Ekdahl, D. Norberg, A.A. Bengtsson, G. Sturfelt, U.R. Nilsson, B. Nilsson, Use of serum or buffer-changed EDTA-plasma in a rapid, inexpensive, and easy-to-perform hemolytic complement assay for differential diagnosis of systemic lupus erythematosus and monitoring of patients with the disease, Clin. Vaccine Immunol. 14 (2007) 549−555.
[27] M. Stegert, M. Bock, M. Trendelenburg, Clinical presentation of human c1q deficiency: How much of a lupus? Mol. Immunol. 67 (2015) 3−11.
[28] P.R. Taylor, A. Carugati, V.A. Fadok, H.T. Cook, M. Andrews, M.C. Carroll, et al., A hierarchical role for classical pathway complement proteins in the clearance of apoptotic cells in vivo, J. Exp. Med. 192 (2000) 359−366.
[29] A.M. Orbai, L. Truedsson, G. Sturfelt, O. Nived, H. Fang, G.S. Alarcon, et al., Anti-C1q antibodies in systemic lupus erythematosus, Lupus. 24 (2015) 42−49.
[30] G. Moroni, S. Quaglini, A. Radice, B. Trezzi, F. Raffiotta, P. Messa, et al., The value of a panel of autoantibodies for predicting the activity of lupus nephritis at time of renal biopsy, J. Immunol. Res. 2015 (2015) 106904.
[31] V.V. Vasilev, R. Noe, M.A. Dragon-Durey, S. Chauvet, V.J. Lazarov, B.P. Deliyska, et al., Functional characterization of autoantibodies against complement component C3 in patients with lupus nephritis, J. Biol. Chem. 290 (2015) 25343−25355.
[32] K. Bork, Pasteurized and nanofiltered, plasma-derived C1 esterase inhibitor concentrate for the treatment of hereditary angioedema, Immunotherapy. 6 (2014) 533−551.
[33] M. Cicardi, W. Aberer, A. Banerji, M. Bas, J.A. Bernstein, K. Bork, et al., Classification, diagnosis, and approach to treatment for angioedema: consensus report from the hereditary angioedema international working group, Allergy. 69 (2014) 602−616.
[34] M. Cicardi, A. Zanichelli, Acquired angioedema, Allergy Asthma Clin. Immunol. 6 (2010) 14.
[35] M.M. Gompels, R.J. Lock, J.E. Morgan, J. Osborne, A. Brown, P. F. Virgo, A multicentre evaluation of the diagnostic efficiency of serological investigations for C1 inhibitor deficiency, J. Clin. Pathol. 55 (2002) 145−147.
[36] M.M. Gompels, R.J. Lock, M. Abinun, C.A. Bethune, G. Davies, C. Grattan, et al., C1 inhibitor deficiency: consensus document, Clin. Exp. Immunol. 139 (2005) 379−394.
[37] S.N. Markovic, D.J. Inwards, R.P. Phyliky, Acquired C1 esterase inhibitor deficiency, Ann. Intern. Med. 133 (2000) 839.
[38] M. Noris, J. Caprioli, E. Bresin, C. Mossali, G. Pianetti, S. Gamba, et al., Relative role of genetic complement abnormalities in sporadic and familial ahus and their impact on clinical phenotype, Clin. J. Am. Soc. Nephrol. 5 (2010) 1844−1859.
[39] C.M. Taylor, S. Machin, S.J. Wigmore, T.H. Goodship, Clinical practice guidelines for the management of atypical haemolytic uraemic syndrome in the united kingdom, Br. J. Haematol. 148 (2010) 37−47.
[40] M. Sridharan, R.S. Go, R.S. Abraham, F.C. Fervenza, S. Sethi, S. C. Bryant, et al., Diagnostic utility of complement serology for atypical hemolytic uremic syndrome, Mayo Clin. Proc. 93 (2018) 1351−1362.
[41] M.R. Tsai, Y.H. Cheng, J.S. Chen, Y.S. Sheen, Y.H. Liao, C.K. Sun, Differential diagnosis of nonmelanoma pigmented skin lesions based on harmonic generation microscopy, J. Biomed. Opt. 19 (2014) 36001.
[42] Z. Prohaszka, L. Varga, G. Fust, The use of 'real-time' complement analysis to differentiate atypical haemolytic uraemic syndrome from other forms of thrombotic microangiopathies, Br. J. Haematol. 158 (2012) 424−425.
[43] S.R. Cataland, V.M. Holers, S. Geyer, S. Yang, H.M. Wu, Biomarkers of terminal complement activation confirm the diagnosis of ahus and differentiate ahus from TTP, Blood. 123 (2014) 3733−3738.
[44] B. Gulbis, A. Eleftheriou, M. Angastiniotis, S. Ball, J. Surralles, M. Castella, et al., Epidemiology of rare anaemias in Europe, Adv. Exp. Med. Biol. 686 (2010) 375−396.
[45] E.C. Miller, N.M. Chase, P. Densen, M.K. Hintermeyer, J.T. Casper, J.P. Atkinson, Autoantibody stabilization of the classical pathway C3 convertase leading to C3 deficiency and neisserial sepsis: C4 nephritic factor revisited, Clin. Immunol. 145 (2012) 241−250.
[46] Y. Levo, Hepatitis B virus and essential mixed cryoglobulinemia, Ann. Intern. Med. 94 (1981) 282.
[47] A. Suarez, E. Vallina, C.A. Navascues, M. Rodriguez, L. Otero, N.G. Sotorrio, et al., Mixed type-II cryoglobulinemia associated with a chronic hepatitis C virus infection, Rev. Clin. Esp. 192 (1993) 325−326.
[48] P.D. Gorevic, Rheumatoid factor, complement, and mixed cryoglobulinemia, Clin. Dev. Immunol. 2012 (2012) 439018.
[49] A.L. Coleman, F. Yu, Eye-related medicare costs for patients with age-related macular degeneration from 1995 to 1999, Ophthalmology. 115 (2008) 18−25.
[50] A.O. Edwards, R. Ritter 3rd, K.J. Abel, A. Manning, C. Panhuysen, L.A. Farrer, Complement factor H polymorphism and age-related macular degeneration, Science. 308 (2005) 421−424.
[51] G.S. Hageman, D.H. Anderson, L.V. Johnson, L.S. Hancox, A.J. Taiber, L.I. Hardisty, et al., A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration, Proc. Natl. Acad. Sci. U S A. 102 (2005) 7227−7232.
[52] J.L. Haines, M.A. Hauser, S. Schmidt, W.K. Scott, L.M. Olson, P. Gallins, et al., Complement factor H variant increases the risk of age-related macular degeneration, Science. 308 (2005) 419−421.
[53] S. McHarg, S.J. Clark, A.J. Day, P.N. Bishop, Age-related macular degeneration and the role of the complement system, Mol. Immunol. 67 (2015) 43−50.

[54] E.C. Schramm, S.J. Clark, M.P. Triebwasser, S. Raychaudhuri, J. M. Seddon, J.P. Atkinson, Genetic variants in the complement system predisposing to age-related macular degeneration: a review, Mol. Immunol. 61 (2014) 118–125.

[55] D. Kavanagh, Y. Yu, E.C. Schramm, M. Triebwasser, E.K. Wagner, S. Raychaudhuri, et al., Rare genetic variants in the CFI gene are associated with advanced age-related macular degeneration and commonly result in reduced serum factor I levels, Hum. Mol. Genet. 24 (2015) 3861–3870.

[56] M.P. Triebwasser, E.D. Roberson, Y. Yu, E.C. Schramm, E.K. Wagner, S. Raychaudhuri, et al., Rare variants in the functional domains of complement factor H are associated with age-related macular degeneration, Invest. Ophthalmol. Vis. Sci. 56 (2015) 6873–6878.

[57] D.H. Anderson, R.F. Mullins, G.S. Hageman, L.V. Johnson, A role for local inflammation in the formation of drusen in the aging eye, Am. J. Ophthalmol. 134 (2002) 411–431.

[58] J. Ba, R.S. Peng, D. Xu, Y.H. Li, H. Shi, Q. Wang, et al., Intravitreal anti-VEGF injections for treating wet age-related macular degeneration: a systematic review and meta-analysis, Drug. Des. Dev. Ther. 9 (2015) 5397–5405.

[59] V.S. Dedania, S.J. Bakri, Sustained elevation of intraocular pressure after intravitreal anti-VEGF agents: what is the evidence? Retina. 35 (2015) 841–858.

[60] J. Hanus, F. Zhao, S. Wang, Current therapeutic development for atrophic age-related macular degeneration, Br. J. Ophthalmol. (2015).

[61] M.J. Tolentino, A. Dennrick, E. John, M.S. Tolentino, Drugs in Phase II clinical trials for the treatment of age-related macular degeneration, Expert. Opin. Investig. Drugs. 24 (2015) 183–199.

[62] T.C. Thomas, S.A. Rollins, R.P. Rother, M.A. Giannoni, S.L. Hartman, E.A. Elliott, et al., Inhibition of complement activity by humanized anti-C5 antibody and single-chain FV, Mol. Immunol. 33 (1996) 1389–1401.

[63] R.P. Rother, S.A. Rollins, C.F. Mojcik, R.A. Brodsky, L. Bell, Discovery and development of the complement inhibitor eculizumab for the treatment of paroxysmal nocturnal hemoglobinuria, Nat. Biotechnol. 25 (2007) 1256–1264.

[64] E.S. Reis, D.C. Mastellos, D. Yancopoulou, A.M. Risitano, D. Ricklin, J.D. Lambris, Applying complement therapeutics to rare diseases, Clin. Immunol. 161 (2015) 225–240.

[65] T. Rondelli, A.M. Risitano, R. Peffault de Latour, M. Sica, B. Peruzzi, P. Ricci, et al., Polymorphism of the complement receptor 1 gene correlates with the hematologic response to eculizumab in patients with paroxysmal nocturnal hemoglobinuria, Haematologica. 99 (2014) 262–266.

[66] J. Zuber, M. Le Quintrec, S. Krid, C. Bertoye, V. Gueutin, A. Lahoche, et al., Eculizumab for atypical hemolytic uremic syndrome recurrence in renal transplantation, Am. J. Transpl. 12 (2012) 3337–3354.

[67] M.A.V. Willrich, B.D. Andreguetto, M. Sridharan, F.C. Fervenza, L.J. Tostrud, P.M. Ladwig, et al., The impact of eculizumab on routine complement assays, J. Immunol. Methods. 460 (2018) 63–71.

Self-Assessment Questions

1. The Classical Pathway recognition molecule is
   a. C1 qrs complex [right answer]
   b. C1 inhibitor
   c. Mannose binding lectin
   d. Pathogen associated molecular patterns
2. The alternative pathway is triggered by
   a. C1 qrs
   b. C3 spontaneous hydrolysis
   c. Ficolins
   d. Factor B
3. The classical pathway is dependent on two metals to activate the cascade and proceed until the formation of the terminal complement complex. Which are they?
   a. Aluminum and lead
   b. Sodium and Potassium
   c. Calcium and Magnesium
   d. Calcium and Potassium
4. In order to measure the classical pathway complement function or activity, the best specimen type is:
   a. EDTA plasma
   b. Citrate plasma
   c. Serum
   d. Urine
5. In order to measure complement activation fragments in circulation, the best specimen type is:
   a. EDTA plasma
   b. Citrate plasma
   c. Serum
   d. Urine
6. What is the immunoglobulin that activates the complement system most efficiently? And what is the one that does not activate complement, respectively?
   a. IgG and IgA
   b. IgD and IgM
   c. IgM and IgG1
   d. IgM and IgG4
7. What is the complement defect associated with hereditary angioedema?
   a. C1 qrs
   b. C1 inhibitor
   c. C4 concentration
   d. C2 concentration
8. The most common acquired autoantibody to complement components, present in atypical hemolytic syndrome and C3 glomerulopathies is:
   a. Antibodies to Factor B
   b. Antibodies to Factor H
   c. C4 nephritic factors
   d. Antibodies to C1q
9. Eculizumab is a monoclonal antibody therapy that targets and blocks a component of the complement cascade. Which is it?
   a. C3
   b. C5
   c. C5b-9
   d. C4

Answers

1. a
2. b
3. c
4. c
5. a
6. d
7. b
8. b
9. b

Chapter 24

Hemoglobin variant detection

Dina N. Greene[1], Ashley Eckel[2], and Christina M. Lockwood[2]

[1]*Washington Kaiser Permanente, Renton, WA, United States,* [2]*Department of Laboratory Medicine, University of Washington, Seattle, WA, United States*

Learning objectives

After reading this chapter, the reader will be able to:

- Describe pathological and nonpathological hemoglobin physiology.
- List biochemical and molecular methods used in hemoglobin analysis.
- Compare quantitative and qualitative methods to detect hemoglobin variants.
- Recognize differences between targeted and comprehensive molecular assays.
- Formulate problem-solving skills to assist in evaluating difficult hemoglobin cases.

Introduction

Hemoglobin (Hb) is the iron-containing protein within red blood cells (RBCs; erythrocytes), responsible for oxygen transport and delivery. In most vertebrates, functional Hb is a tetramer, consisting of four globular protein subunits organized as two heterodimers (Fig. 24.1 and Table 24.1). Each heterodimer is composed of one alpha or alpha-like globin subunit and one beta or beta-like globin subunit. Each of the four subunits surrounds a central heme group that contains iron and binds one oxygen molecule. Hb is often recognized for its cooperative binding; as each individual Hb monomer reversibly binds to one molecule of oxygen, the shape of the tetramer shifts, increasing the affinity for oxygen binding by neighboring subunits.

There are three normal Hbs detected in systemic circulation: adult hemoglobin (HbA), fetal hemoglobin (HbF), and HbA_2 (Table 24.1). All three of these Hb tetramers have two alpha globin subunits. The HbA tetramer consists of two alpha and two beta subunits, which are encoded by two alpha globin genes (*HBA1* and *HBA2*) and one beta globin gene (*HBB*), respectively. The "A" indicates adult, and HbA comprises >90% of normal Hb in children and adults. In contrast, in HbF, the alpha subunits pair with two gamma globin subunits, beta-like globins encoded by two gamma globin genes (*HBG1* and *HBG2*). The "F" indicates fetal, and HbF is the primary species of Hb present in utero (after the first trimester; Fig. 24.2) and at birth. HbA_2 is a minor Hb species, built from two alpha and two delta globin subunits, which appears at birth and is consistently present in normal individuals at a relative concentration of ~2.0%–3.5% of total Hb. The ratios of these three types of Hb are altered in certain pathophysiological states, including sickle cell anemia (increased HbF) and certain thalassemias (increased HbA_2).

The process by which the proportions of the major Hb subunits change over time in an individual is referred to as developmental switching (Fig. 24.2). During embryogenesis, there are primitive forms of Hb, including zeta globin (alpha-like) and epsilon globin (beta-like) subunits. Both of these globins are important during gestation in order to support the oxygen requirements of an embryo. After approximately 9 weeks of development, embryogenesis is complete and the Hb species transition into those that are required to support the increased oxygen demands of the fetus. During this phase of gestation, HbF becomes

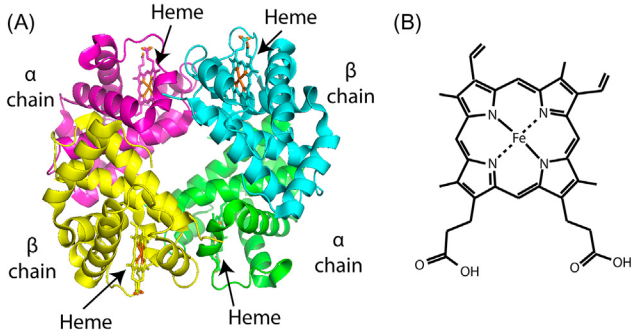


FIGURE 24.1 Hemoglobin structure. (A) Functional hemoglobin is a tetramer, consisting of four globular protein subunits organized as two heterodimers. (B) Each subunit surrounds a central heme group, which contains an iron and binds one oxygen molecule. *Image in (A) produced using pymol, hemoglobin 2hhb.*

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00024-7

TABLE 24.1 Hemoglobin tetramers detected in normal profiles.

	Subunit composition	Genes encoding subunits	Normal relative percent	Elevated in	Decreased in
HbA	$\alpha\alpha\beta\beta$	*HBA1*	96	n/a	α or β point mutations
		HBA2			
		HBB			
HbF	$\alpha\alpha\gamma\gamma$	*HBA1*	1	Hereditary persistence of fetal hemoglobin; sickle cell	n/a
		HBA2			
		HBG1			
		HBG2			
HbA_2	$\alpha\alpha\delta\delta$	*HBA1*	3	β-Thalassemia	α-Thalassemia
		HBA2			
		HBD			

HbA, Adult hemoglobin; *HbF*, fetal hemoglobin.

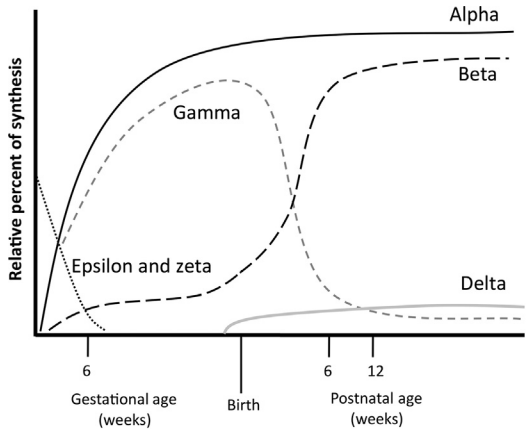


FIGURE 24.2 Developmental switching is the process by which hemoglobin species appear and diminish in response to the changing physiological needs of human development.

the primary form of Hb found in fetal circulation. At birth, HbF rapidly decreases and is replaced with HbA. In adulthood, HbF normally persists as approximately 1%–4% of total adult Hb.

Around 5%–7% of the world's population carries one or more clinically important mutations or deletions affecting the globin genes (Table 24.2). It is, however, also important to note that the Hb genes are highly polymorphic and rare structural variants, which are collectively frequent, can be clinically insignificant. Pathological Hb abnormalities are broadly categorized into two forms: structural hemoglobinopathies and thalassemias. Structural hemoglobinopathies occur when there is a mutation, leading to a structurally abnormal Hb subunit. These hemoglobinopathies are most commonly caused by point mutations that lead to an amino acid change in the primary protein sequence. In the United States, sickle cell disease, a structural hemoglobinopathy caused by mutations in the beta globin gene, is the most common inherited blood disorder and leads to production of a variant Hb termed HbS. Approximately 10% of African Americans carry an HbS allele. In HbS, a missense mutation causes the substitution of a valine for glutamic acid in the sixth amino acid (Glu6Val) of the beta globin gene *HBB*. The mutated HbS polymers reduce red cell deformability and distort erythrocytes into the characteristic sickle-shaped cells that can occlude blood vessels, particularly in patients homozygous for the mutation (HbSS) (Fig. 24.3). The clinical significance of various hemoglobinopathies varies by the specific abnormalities. For example, patients with HbC disease have a variant Hb caused by a mutation in the same amino acid of the beta globin gene as HbS. However, in HbC, there is substitution of lysine for the glutamic acid at the sixth amino acid (Glu6Lys) of the beta globin gene, and this change is associated with less clinical severity than sickle cell disease. The hemoglobinopathies are generally associated with a spectrum of phenotypic severity depending on how many gene mutations are involved. For example, heterozygous hemoglobinopathy mutations do not typically have immediate clinical symptoms, but homozygous mutations [sickle cell disease (HbSS)] compound heterozygous mutations (HbSC disease), or hemoglobinopathies with a coincident thalassemia can manifest with a range of clinical phenotypes.

Thalassemias, in contrast to the structural hemoglobinopathies, are defined by a quantitative decrease in the

TABLE 24.2 Commonly detected abnormal hemoglobin variants.

	Subunit composition	Clinical indication
HbS	αα$\beta^{*}\beta^{*}$ *Indicates Glu6Val mutation	Not pathological in heterozygous cases; homozygous indicates sickle cell disease
HbC	αα$\beta^{*}\beta^{*}$ *Indicates Glu6Lys mutation	Not pathological in heterozygous cases; often occurs as a compound heterozygous genotype with HbS; homozygous shows a mild-to-moderate anemia
HbE	αα$\beta^{*}\beta^{*}$ *Indicates Glu26Lys mutation	Not pathological in heterozygous cases; considered a thalassemic hemoglobinopathy; homozygous shows a mild-to-moderate anemia
HbH	β_4	Detected only when two or more α genes are deleted
HbBarts	γ_4	Specific to pediatric patients; detected only when two or more α genes are deleted

For more details, see Refs. [1–3].

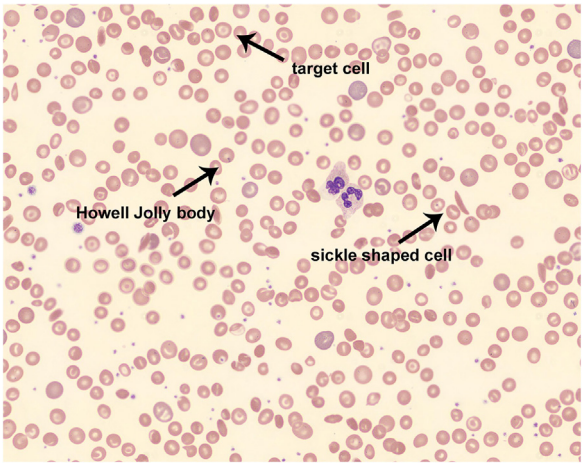


FIGURE 24.3 Peripheral smear from a patient with sickle cell disease illustrates the spectrum of red blood cell findings in this disorder, including sickle cells, target cells, Howell–Jolly bodies, and polychromasia.

overall amount of an Hb subunit. The thalassemias are classified according to the affected globin chain(s). Alpha-thalassemia, for example, occurs when there is a decreased concentration of alpha globin; beta-thalassemia occurs when there is a decreased concentration of beta globin. Thalassemias are most commonly caused by multiexon deletion mutations that can occur due to nonhomologous recombination events, but can also be due to individual point variants that reduce the structural integrity and stability of the transcript or the protein. The latter are considered thalassemic hemoglobinopathies and include Hb Constant Spring and HbE, two hemoglobinopathies more frequently found in individuals of Chinese and Southeast Asian ancestries. Clinically, thalassemias are classified according to the severity of the symptoms as thalassemia minor (asymptomatic), thalassemia intermedia (nontransfusion-dependent anemia), and thalassemia major (transfusion-dependent, severe anemia).

Detection of Hb abnormalities relies heavily on the clinical laboratory. A variety of biochemical and molecular techniques are applied for characterization of hemoglobinopathies. Diagnostic testing for Hb disorders is frequently performed in the evaluation of anemia and microcytosis, pulmonary hypertension, and preoperative anesthesia evaluation, but testing is also increasing in volume due to implementation of prenatal and population-based screening programs. For example, while a heterozygous mutation may be clinically asymptomatic, it can have important consequences for preconception and prenatal screening, particularly in high-risk demographic populations. Prenatal and postnatal testing are also complicated by the fact that, because of developmental switching of Hbs, detection of clinically severe hemoglobinopathies can be delayed until after the first trimester (if the mutations are in the alpha genes) or after birth (if the mutations are in the beta genes). Proper laboratory diagnosis is crucial for characterizing the different forms of hemoglobinopathies with important implications for prevention and treatment.

Blood smears and cell counts

The first indication of a hemoglobinopathy is frequently an abnormality in a complete blood count (CBC). The CBC test incorporates a number of hematologic parameters important for the evaluation of a hemoglobinopathy. The key components of the CBC for detection of hemoglobinopathies are the red cell indices, which include Hb, hematocrit (HCT), RBC count, mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), and red cell distribution width (RDW). A CBC generally also includes a white blood cell (WBC) count and a platelet count.

Hb abnormalities that lead to decreased erythrocyte stability can cause a general anemia, identified in the

laboratory by a decrease in Hb and HCT. A manual differential may show a range of phenotypic erythrocyte abnormalities. Blood smears from patients with sickle cell trait (HbAS) are usually entirely normal, occasionally with a small number of target cells; however, patients with sickle cell anemia (HbSS) can show sickle cells, target cells, anisocytosis, polychromasia, basophilic stippling, and nucleated red cells. A blood film from a patient with Hb C disease may show a large number of target cells, microcytosis, and characteristic Hb C crystals. In the thalassemias, the reduced production of Hb also leads to a uniformly reduced cell volume called microcytosis (low MCV) with hypochromasia (reduced MCH). The RBC count can be increased as a compensatory mechanism. In these patients, the RDW, which measures the degree of variation in erythrocyte size, often remains normal. The blood films from a patient with thalassemia trait may show target cells, elliptocytes, and basophilic stippling.

In assessing patients with anemia and a low MCV, it is critical to consider the potential for iron deficiency, also characterized by anemia and low MCV but frequently with a low RBC count and an increased RDW. Testing specific for assessing an individual's iron status should be performed and may include ferritin, transferrin, and total iron binding capacity. Thalassemias can be distinguished from iron deficiency anemia by the RBC count (usually increased in thalassemia and decreased in iron deficiency) and red cell width (usually normal in thalassemia and decreased in iron deficiency).

The RBC findings are critical to the characterization of a hemoglobinopathy, and a CBC should always be integrated into the testing strategy for hemoglobinopathies, regardless of the other tests used. While CBC findings alone are not diagnostic of Hb abnormalities, the CBC findings play a critical role in their ability to help the provider assess the severity of a hemoglobinopathy.

Case Example 1

A 22-year-old African American male who was diagnosed at birth with sickle cell disease presents to his primary care physician with complaints of fatigue and unbearable leg muscle pain. His CBC results are listed in Table 24.3.

Are these results expected for an individual with HbSS (homozygous genotype)?

TABLE 24.3 Laboratory results for Case Example 1.

Hematologic parameter (units)	Result	Reference interval
WBC ($10^3/\mu L$)	11.2	4.3–10.0
RBC ($10^6/\mu L$)	6.0	4.4–5.6
Hb (g/dL)	9.8	13.0–18.0
HCT (%)	30	38–50
MCV (fL)	99	81–98
MCH (pg)	27.5	27.3–33.6
MCHC (g/dL)	34.0	32.2–36.5
Platelet count ($10^3/\mu L$)	220	150–400

Complete blood count remarkable for elevated white cell count, red cell count, and mean corpuscular volume with accompanying decrease in hemoglobin and hematocrit.
Hb, Hemoglobin; *HCT*, hematocrit; *MCH*, mean corpuscular hemoglobin; *MCV*, mean corpuscular volume; *RBC*, red blood cell; *WBC*, white cell count.

Case 1 Resolution

HbS is defined by the substitution of a valine for glutamic acid in the sixth amino acid of the beta globin protein (Table 24.2). Sickle cell disease occurs when both genes have this point mutation and the pathological consequences are significant. HbS has a propensity to polymerize within the erythrocyte, which decreases the flexibility of the membrane and leads to a sickling shape. The altered conformation of the erythrocyte is less flexible and therefore does not retain the mobility necessary to traverse the vasculature. Overall, this results in vaso-occlusive episodes (sometimes called sickle cell crises) and hemolytic anemia.

The CBC results described here are expected for a patient with sickle cell disease. Hb and HCT are decreased as a result of the anemia; WBCs are increased from inflammation; MCV and RBC count are increased as physiological attempts to compensate for the anemia. Palliative care should be used to help relieve this patient of their symptoms. The CBC findings are likely not severe enough to warrant blood transfusion; however, the decision when to give transfusions to these patients can depend on the patient's subjective symptoms and nonlaboratory-based vital sign measurements (i.e., heart rate and blood pressure).

Biochemical techniques

Quantitative measurement of hemoglobin variants

The relative concentrations of the various reference and Hb variants have clinical relevance; therefore accurate quantification of Hb species assists in the proper diagnosis of these Hb abnormalities. A normal Hb profile has a relative globin ratio of approximately 96% HbA, 1% HbF, and 3% HbA_2 (Fig. 24.4 and Table 24.1).

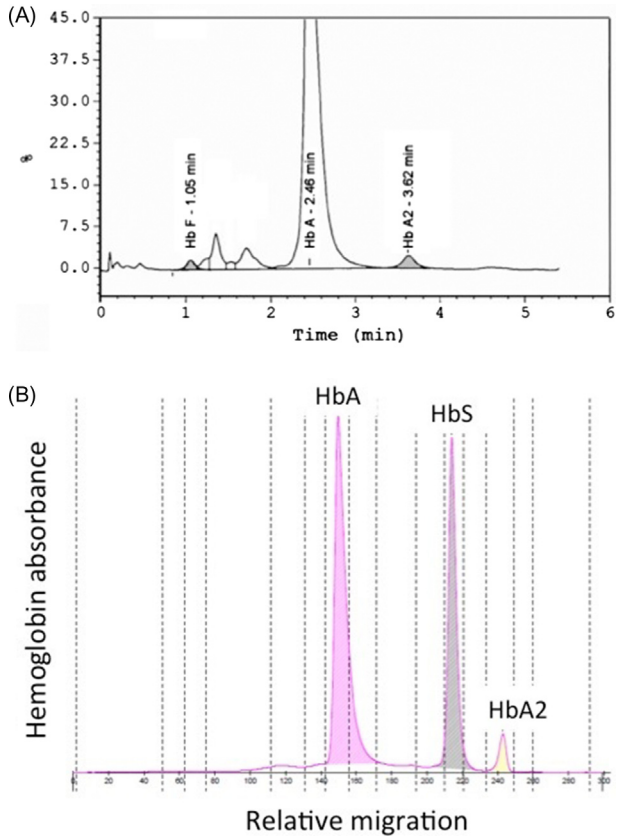


FIGURE 24.4 Representative high-performance liquid chromatogram from an adult with no detectable hemoglobin variants (A) and capillary zone electropherogram from an HbS carrier (B).

Abnormalities in the Hb profile are defined either as a disruption in this ratio or as a visual indication of a quantifiable species not observed in normal individuals.

High-performance liquid chromatography (HPLC) and capillary zone electrophoresis (CZE; sometimes referred to as CE) are the two primary methods used by clinical laboratories to measure Hb species. Both techniques require small amounts of sample and can be performed by automated methods, making them amenable to high-or low-volume laboratory workflows. Some laboratories employ one or more methods in their protocol for detection of hemoglobinopathies. Choice of method depends on the function of the laboratory and its testing algorithm. It is important to note that both HPLC and CZE are generally considered screening methodologies; therefore separate techniques, including alternative electrophoresis methods and molecular testing methods, are often used as confirmatory tests. Depending on the patient population served by a clinical laboratory and the indication for testing, laboratories may have different algorithms for testing. For example, some laboratories employ a screening method (HPLC or CZE) and only perform confirmatory testing on samples that screen positive, while other laboratories may evaluate all samples by multiple methods without requiring a positive screening result. The latter workflow is advantageous for high-risk populations, because certain alpha-thalassemias are not distinguishable from normal individuals by the screening methods alone (HPLC or CZE).

High-performance liquid chromatography

The classical method used by clinical laboratories to separate and quantify Hb variants is cation exchange HPLC (Fig. 24.4A). In this method, a small amount of red cell lysate is injected, and positively charged Hb is retained on a negatively charged silica-based column. Hb species are eluted from the column using a liquid mobile phase with phosphate buffers that have a gradient of salt concentrations and pH to decrease the affinity of the Hbs to the column. The eluted Hb is detected with absorbance wavelengths for maximal Hb detection (415 nm). Retention time, or the time it takes to elute from the column, is used to distinguish the major Hb species. The areas under the individual chromatographic peaks are integrated and compared to the total area under all peaks to quantify the corresponding percent Hb concentration. HPLC runs are preceded by priming and calibration of the instrument, which involves calibration factors containing HbA_2 and HbF that must measure at expected values.

HPLC allows quantification of the three normal Hbs, including HbF and HbA_2, as well as a large number of other Hb variants. Common and rare Hb variant profiles using HPLC are well defined, and libraries of chromatograms have been published to facilitate interpretation. Interpretation can be complicated by the biological background "noise" caused by posttranslationally modified and degraded Hb species. These modifications can produce small or shifted peaks, sometimes coeluting with other Hbs. For example, glycated HbS can elute in the HbA window. In addition, since HbE coelutes with HbA_2, the two cannot be differentiated using this technique. HbH and Bart Hb, which are clinically significant Hbs causing alpha-thalassemia, may elute too quickly from a column depending on the specific methods used and preclude their quantitation. Interestingly, many laboratories use HPLC-based assays to quantify Hb A1c for the management of diabetes mellitus. The presence of variant Hbs can be observed when the assay for Hb A1c is performed, and is sometimes the first indication a patient carries a variant Hb.

Capillary zone electrophoresis

CZE is a relatively recent technique compared with HPLC for the detection of abnormal Hbs, but its use by the clinical laboratory has significantly increased over the last decade. Hb separation and quantification is

particularly amenable to this technique. In CZE, separation is based on movement of charged molecules (the Hb fractions) in a silica-based capillary tube in an alkaline buffer, containing electrolytes that create an electroosmotic flow. Hb absorbance is monitored at the base of the capillary to detect migrating Hb species. The resulting peaks are divided into zones that are defined by their migration relative to HbA; the threshold position for HbA is established at the beginning of each run using control material and must be determined for each capillary. Thus samples must contain some amount of HbA to characterize the migration pattern of the other Hb components. Samples without endogenous HbA, including homozygous beta globin mutations and fetal samples, must be mixed with a reference sample, containing HbA to establish and identify the migration pattern. The results from the mixed and unmixed samples are compared to identify the unknown peaks and to determine the relative Hb concentrations in the sample lacking HbA. The data output from CZE is called an electropherogram, and, unlike HPLC, there is generally very little background noise to distinguish from relevant Hb species (Fig. 24.4B). CZE also allows clean separation of HbE from HbA_2 and facilitates easier detection of HbH and HbBart.

Case Example 2

A 6-month-old female child presented with progressive pallor and failure to thrive. On examination, she had significant hepatosplenomegaly. A CBC showed a severe anemia (Table 24.4) with marked anisopoikilocytosis (variable cell size and shape), and she was referred for further investigations. Hb HPLC showed a markedly increased HbF and absent HbA. Both parents were found to have raised HbA_2 fractions on HPLC around 6% (normal range 2.0%–3.5%; Fig. 24.5). What is the most likely diagnosis in the child? Is this the likely cause of the child's anemia? What is the most likely cause of the elevated HbA_2 in the parents?

TABLE 24.4 Laboratory results for Case Example 2.

Hematologic parameter (units)	Result	Reference interval
WBC ($10^3/\mu L$)	15.2	4.3–10.0
RBC ($10^6/\mu L$)	4.3	4.4–5.6
Hb (g/dL)	8.9	13.0–18.0
HCT (%)	32	38–50
MCV (fL)	74	81–98
MCH (pg)	29.1	27.3–33.6
MCHC (g/dL)	33.6	32.2–36.5
Platelet count ($10^3/\mu L$)	220	150–400

Complete blood count remarkable for elevated white cell count with accompanying decrease in red cell count, mean corpuscular volume, hemoglobin, and hematocrit.
Hb, Hemoglobin; *HCT*, hematocrit; *MCH*, mean corpuscular hemoglobin; *MCV*, mean corpuscular volume; *RBC*, red blood cell; *WBC*, white cell count.

Case 2 Resolution

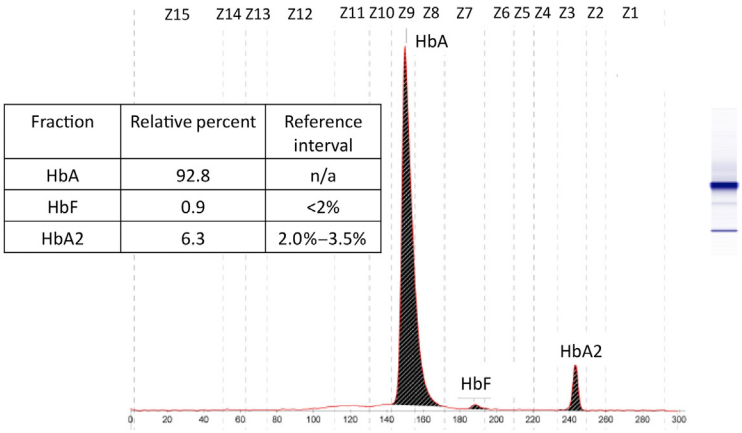
HbA_2 is a tetramer of two alpha and two delta globin subunits. Elevated HbA_2, as seen in the parents of the child, occurs in beta-thalassemia due to decreased production of beta globin. This causes a molar excess of free alpha globin subunits available to bind delta globin. The net result is an elevated relative percentage of HbA_2. The findings in this case suggest single-gene beta globin deletions in each of the parents. Single-gene beta globin deletions are not pathological, but they are important to diagnose as prenatal precautions. More severe forms of beta-thalassemia, as seen in this child, are referred to as beta-thalassemia major and occur when both beta globin genes have deleterious mutations that result in the absence of or significant decrease in beta globin synthesis. Rather than elevated HbA_2, the most striking Hb profile abnormality in beta-thalassemia major is a significant decrease in or complete absence of HbA with no accompanying variant peak. A chronic hemolytic anemia occurs in beta-thalassemia major, leading to blood transfusion dependence, which often results in iron overload. Resulting multi-organ damage is common.

Case Example 3

A 7-year-old African American female was brought to her pediatrician because her parents have noticed that she becomes out of breath when playing with peers and frequently complains of muscle cramps. Her records indicate that she was diagnosed as a carrier for HbS via newborn screening shortly after birth. Her pediatrician ordered a CBC and Hb profile evaluation (Table 24.5). Are the results consistent with the patient's diagnosis as a carrier for HbS?

Case 3 Resolution

In HbS carriers, the HbA:HbS ratio is approximately 60:40 (Table 24.6). Distinguishing between HbS trait and HbS with a concurrent beta-thalassemia can be difficult in the neonatal and early postnatal periods because the elevated concentration of HbF can mask the expected ratio between HbA and HbS observed in adults. If there is an underlying beta-thalassemia, the ratio decreases and the relative concentration of HbS is greater than that of HbA. Reviewing the relative percentages of Hb species in this child shows that her HbA is 28%, while her HbS is



Fraction	Relative percent	Reference interval
HbA	92.8	n/a
HbF	0.9	<2%
HbA2	6.3	2.0%–3.5%

FIGURE 24.5 Capillary zone electropherogram shows an elevated HbA_2, which is indicative of beta-thalassemia. Migration identification zones are indicated with "Z" followed by a number.

TABLE 24.5 Laboratory results for Case Example 3.

Hematologic parameter (units)	Result	Reference interval
WBC ($10^3/\mu L$)	13.2	4.5–13.5
RBC ($10^6/\mu L$)	5.7	4.0–5.2
Hb (g/dL)	8.6	11.5–15.5
HCT (%)	29	35–45
MCV (fL)	99	77–95
MCH (pg)	26.1	25.0–33.0
MCHC (g/dL)	34.9	32.2–36.5
Platelet count ($10^3/\mu L$)	358	200–450
HbA (%)	28.0	n/a
HbF (%)	7.0	<2
HbA_2 (%)	5.2	2.0–3.5
HbS (%)	58.8	0
IEF interpretation	HbS and HbA present	

Complete blood count remarkable for elevated red cell count and mean corpuscular volume with accompanying decrease in hemoglobin and hematocrit. HbS was detectable by biochemical methods, and HbA_2 was elevated.
Hb, Hemoglobin; *HbA*, adult hemoglobin; *HbF*, fetal hemoglobin; *HCT*, hematocrit; *IEF*, isoelectric focusing electrophoresis; *MCH*, mean corpuscular hemoglobin; *MCV*, mean corpuscular volume; *RBC*, red blood cell; *WBC*, white cell count.

58.8%; HbA_2 and HbF are also elevated (5.2% and 7.0%, respectively; Table 24.5). Paternal and maternal Hb profiles are reviewed to help clarify the lineage: mother was a known HbS carrier; father had previously screened normal. On retesting, the father was shown to have an elevated HbA_2 (4.8%). Molecular testing confirmed that this child inherited one mutated beta globin gene from each parent, one with a point mutation conferring HbS and the second with a promoter mutation conferring a beta-thalassemia. In patients with this genotype, the clinical severity is inversely proportional to the concentration of HbA (higher percent HbA correlates with less severe presentation). Events that stress the body (illness and extended physical activity) can push the body into crisis where transfusion and/or hospitalization are necessary.

Qualitative measurement of hemoglobin variants

In addition to those Hb variants above, a large number of less common variants have been described in various ethnic populations (several are listed in Table 24.7). A conceptually straightforward approach to Hb variant detection is asking the question, "Does a simple biophysical challenge or basic separation technique qualitatively show anything different about an individual's results that may reflect their endogenous Hb?" These assays have historically been fundamental for intersecting Hb abnormalities with the clinical laboratory. Importantly, many of these techniques still play a large role in the laboratory assessment of Hb.

Slab gel electrophoresis

There are three types of electrophoresis that are commonly used to separate and visualize Hb species: acid, alkaline, and isoelectric focusing. Each of these techniques applies a small amount of lysed red cells to a polymer matrix for separation using an electric charge. The process is largely manual and semiquantitative; scanning densitometry is used to assess the Hbs present and lacks the precision to provide the attenuated details offered by HPLC and CZE. However, slab gel electrophoresis methods are useful confirmatory techniques for abnormalities detected using HPLC or CZE. In addition, isoelectric focusing electrophoresis (IEF) offers clinical information unable to be resolved using aforementioned biochemical methods.

TABLE 24.6 Distinguishing features between phenotypes remarkable for HbS.

Hemoglobin profile	Clinical severity	Hematocrit and hemoglobin	Red cell morphology	Hemoglobin profile quantification
AS	Normal	Normal	Rare sickle cell	HbA 60%; HbS 40%
SS	Marked	Low	Target cells; 20%–30% sickle cells	HbS 80%–95%; HbF 2%–20%
S/β^0-Thalassemia	Moderate to marked	Low	Hypochromic; microcytic; target cells; rare sickle cells	HbS 80%–95%; HbF 1%–15% F; HbA_2 3%–6%
S/β^+-Thalassemia	Mild to moderate	Low/normal	Slightly microcytic; target cells; rare sickle cells	HbS 55%–75%; HbA 10%–30%; HbF 1%–10%; HbA_2 3%–6%
SC	Mild to moderate	Low/normal	Slightly microcytic; rare sickle cells	HbS ~50%; HbC ~50%

β^0-Thalassemia indicates a complete gene deletion in a single beta gene; β^+-thalassemia indicates a mutation that decreases expression of a single beta gene.
HbA, Adult hemoglobin; *HbF*, fetal hemoglobin.

TABLE 24.7 Rare hemoglobin variants.

Hemoglobin	Amino acid mutated
Alpha globin gene mutations	
Hb Hasharon	47
G-Philadelphia	68
Q-Thailand	74
Nigeria	81
Constant Spring	142
Beta globin gene mutations	
J-Baltimore	16
N-Baltimore	95
Detroit	95
O-Arab	121
Hb D Los Angeles (D-Punjab)	121
Hb Lepore	Fusion Hb (beta and delta)

Hb electrophoresis exploits the migration changes in Hb species when subjected to various pH environments and an electric field. It can be performed under acidic (pH 6.0–6.2; performed on citrate agar or an appropriate agarose gel) or alkaline (pH 8.0–8.6; generally performed on cellulose acetate) conditions. As the name implies, a red cell hemolysate is applied to wells on one end of a slab of gel, and an electric field is then applied to force the Hbs to migrate across the gel. Depending on whether acid or alkaline conditions are used, and the type of gel, the Hbs become charged, such that their charge-based migration patterns are predictable between the cathodal (negatively charged) and anodal (positively charged) ends of the gel. Visualization of the migrated Hb species is accomplished using a nonspecific protein stain (generally Ponceau S), which is sufficient, since 98% of the protein within a red cell corresponds to Hb, and scanning densitometry. A small, but noticeable band corresponding to carbonic anhydrase will also be detected. Neither acid nor alkaline electrophoresis provides a complete resolution of normal Hb species from common pathological Hb species (Fig. 24.6A and B). For example, while using alkaline electrophoresis, HbA_2 comigrates with HbC and HbE; on acid electrophoresis, HbA comigrates with HbA_2, HbE, HbD, and HbG. Consequently, both types of electrophoresis are functionally equivalent, but can provide complementary data in the investigation of an Hb abnormality. There are commercially available kits to detect Hb using acid or alkaline electrophoresis.

The third type of slab gel electrophoresis, IEF, is considered the biochemical gold standard for Hb abnormality detection and has two main advantages over acid and alkaline electrophoresis. First, because IEF separates proteins based on their molecule-specific isoelectric point (pI), the resolution of various Hb species is more refined than alkaline or acid gel electrophoresis. Second, IEF uses a heme-specific stain such as o-dianisidine rather than a nonspecific protein stain, which enables the technique to be more analytically sensitive and specific. IEF can detect small concentrations of HbH, the beta globin homotetramer that serves as the pathological indicator of alpha-thalassemia. HbH is often undetectable by other biochemical techniques (including CZE and HPLC).

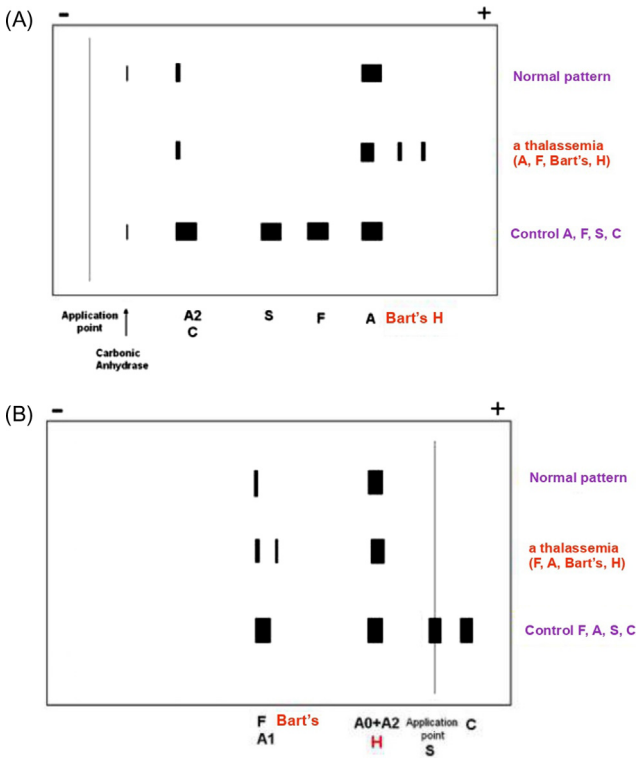


FIGURE 24.6 Diagram of normal and common variant hemoglobin migration in alkaline (A) and acid (B) electrophoresis.

Sickle solubility test

The sickle solubility test is a qualitative screening assay for sickling Hbs. This test exploits the inherent propensity for deoxygenated (deoxy-) HbS to polymerize and precipitate out of solution under conditions where deoxy-HbA remains soluble. In this assay, an aliquot of red cells is lysed in a reducing environment buffered to enhance the insolubility of deoxy-HbS. The appearance of a cloudy, opaque precipitate indicates a positive result. In general, the test will be positive when the concentration of HbS (or another sickling Hb) is greater than approximately 20%, which means that both homozygous and heterozygous carriers will be detected. Although the sickle solubility test cannot differentiate between homozygous and heterozygous genotypes, it is frequently used for screening at smaller volume hospitals, emergency department settings, and for screening blood donors for the presence of HbS.

Unstable hemoglobins

The biophysical properties of certain Hb variants make them more susceptible to oxidation relative to HbA. Assays for unstable Hb exploit this difference by detecting precipitates or inclusions that form when a whole blood lysate is exposed to heat or isopropanol, which enhance oxidation and destabilize the Hb bonds. These

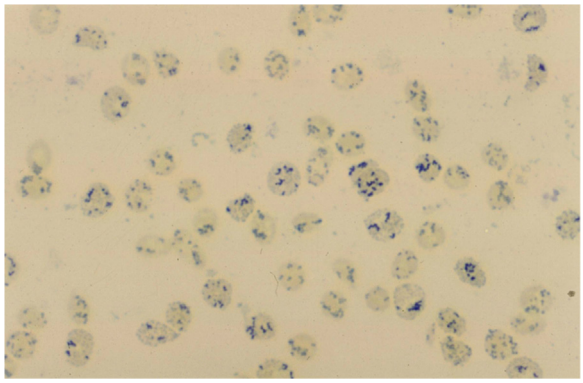


FIGURE 24.7 Red blood cell smear stained with bromocresol blue is remarkable for Heinz bodies. Concurrent biochemical analysis indicates that this individual has HbH disease, a form of alpha-thalassemia.

inclusions are called Heinz bodies and can also be detected by evaluating whole blood smears with supravital stains such as methyl violet and bromocresol blue (Fig. 24.7). Heinz bodies are also seen physiologically in G6PD deficiency, a genetic disorder that causes hemolytic anemia.

Case Example 4

A 28-year-old pregnant Filipino female presents for wellness prenatal screening. Her CBC results indicate a mild microcytic anemia (MCV = 72), which encourages her physician to order an Hb screening panel. The laboratory analyzes the sample using CZE and isoelectric focusing. No abnormalities are detected using CZE, but the IEF demonstrates an HbH concentration of <2% (Fig. 24.8). What is this indicative of?

Case 4 Resolution

Alpha-thalassemia occurs when there is a decreased synthesis of alpha globin, which disrupts the normal molar ratio of alpha to beta globin subunits and leads to a relative excess of beta globin. The free beta globin subunits form homotetramers, called HbH, which are inherently unstable and precipitate to form inclusion bodies within the red cell. The clinical severity of alpha-thalassemia is directly proportional to the number of alpha globin genes deleted. While single-gene deletions are usually silent, two-gene deletions may cause mild microcytic anemia, and three-gene deletions can lead to moderate or severe hemolytic anemia (known as HbH disease). Notably, four-gene deletions cause a condition termed hydrops fetalis in which excess fluid builds up in the fetus before birth in addition to anemia and other organ abnormalities. Most affected fetuses are stillborn or die shortly after birth.

Traditionally, the laboratory has detected HbH disease using biochemical methods because HbH concentrations

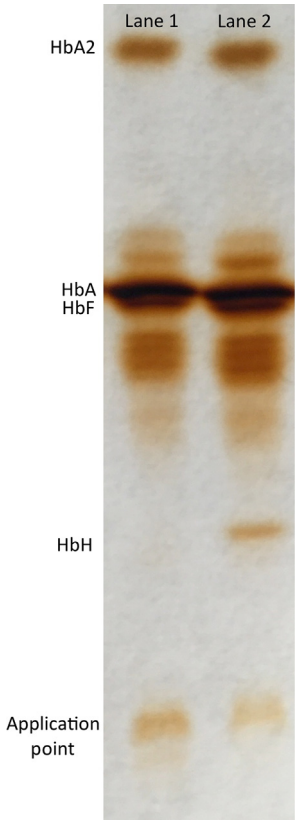


FIGURE 24.8 Isoelectric focusing electrophoresis of normal control (Lane 1) and patient remarkable for HbH of <2% (Lane 2). Major hemoglobin species and application point are appropriately marked; remaining bands are posttranslationally modified and degraded hemoglobin species.

are considerably elevated in these patients (~5%–15%). However, low concentrations of HbH (<2%) may be identified using IEF and indicate two-gene deletions. This patient has an MCV and an HbH concentration consistent with deletion of two alpha globin genes. There is no immediate medical concern for her, but because she is seeking prenatal care, paternal testing should be performed. If the father has evidence of a hemoglobinopathy, particularly one or more alpha globin gene deletions, genetic counseling services should be offered.

Mass spectrometry

Detection of Hb abnormalities using mass spectrometry is a relatively new application that has not been widely adopted in the clinical laboratory. There are various methods for mass spectrometry, but the general principle involves bombarding the sample with a laser beam and assessing the mass-to-charge ratio of produced ions. Mass spectrometry offers enhanced sensitivity and specificity for detection of variant Hbs, but the implementation of this method in the clinical laboratory can be costly and time-consuming.

Currently, Hb profiling by mass spectrometry has been most successful for newborn screening because of the amenability to small sample volumes, including the dried blood spots used in the newborn screen. Mass spectrometry also has invaluable utility for harmonization and standardization efforts. Standardization allows clinical thresholds to be universally adopted. For Hb variant analysis, this is most important for quantification of HbA_2, by which between-method standardization would allow for the endorsement of international beta-thalassemia diagnostic guidelines. The International Federation for Clinical Chemistry (IFCC) has supported a standardization program for HbA_2 that includes a scientific working group. Standardization requires a reference method, a certified reference material, and defined analytical performance expectations. The working group has made successful progress in developing an isotope dilution tandem mass spectrometry assay that quantifies HbA and HbA_2, allowing for the calculation of $\%HbA_2$, which is the recommended unit of measurement. This method meets the working group defined performance specifications and has the ability to assign primary calibrator values that could vastly improve comparability of results.

Molecular techniques

Although some Hb defects can be acquired or occur de novo, most Hb disorders are caused by inherited genetic defects. The majority of Hb disorders are autosomal recessive, and individuals heterozygous for a variant are generally asymptomatic. Molecular methods are often a crucial complement to hematologic and biochemical methods after a presumptive identification of a hemoglobinopathy and/or thalassemia. Laboratory determination of the underlying mutation(s) has utility in establishing a definitive diagnosis in cases where the hematologic phenotype is complex and also has utility in elucidating carrier status for genetic counseling.

There are greater than 1600 Hb variants that have been catalogued on a publicly available globin gene server database (HbVar: http://globin.cse.psu.edu). These variants range from single-nucleotide substitutions to chromosomal deletions and duplications. Thus a variety of molecular techniques must be employed to detect the various types of mutations seen in hemoglobinopathies. Current molecular methods can be separated into targeted assays that are designed to detect recurrent variants and methods suited for the identification of unknown or less common variants. Regardless, many of the methods start with an amplification step by PCR followed by one or more of an assortment of methods available for product detection. Each method has its own advantages and

disadvantages, and the most appropriate laboratory approach depends on both available technical expertise as well as the type and variety of Hb variants likely to be encountered in the relevant patient population.

The vast majority of cases submitted for laboratory analysis are for postnatal detection of variants; therefore DNA can be extracted from peripheral blood leukocytes. For in utero diagnosis, DNA can be extracted from chorionic villi obtained between gestational weeks 10 and 13, amniocytes obtained in the second trimester of pregnancy, or, most recently, from cell-free fetal DNA in maternal circulation.

Targeted assays for known variants

PCR-based techniques used for Hb variant identification include gap-PCR, allele-specific oligonucleotide hybridization or dot-blot analysis, reverse dot-blot analysis, allele-specific PCR, amplification refractory mutation system, high-resolution melt curve analysis, restriction enzyme analysis, and multiplex ligation-dependent probe amplification (MLPA). An efficient strategy for identifying common variants is to simultaneously detect multiple mutations using a multiplex assay. Several of the above methods have been designed as multiplex assays.

Gap-PCR

While more than 100 molecular forms of alpha-thalassemia have been identified, more than 90% of variants are comprised of a common set of deletions (the exact percent varies by population). This has led to the development and widespread adoption of gap-PCR for alpha-thalassemia diagnostics, a method that multiplexes detection of the seven most common alpha globin deletions: −3.7, −4.2, -FIL (Filipino deletion), -THAI (Thailand), -MED (Mediterranean), −20.5, and -SEA (Southeast Asian, Fig. 24.9). As several of the names imply, each deletion was initially described in a specific demographic. However, as global populations are increasingly mobile, there has been an increased incidence of alpha-thalassemia beyond the first-described populations. Gap-PCR can also be utilized to detect beta-thalassemia, delta−beta-thalassemia, hereditary persistence of fetal hemoglobin (HPFH), and the rare gamma−delta−beta-thalassemia.

Gap-PCR can be multiplexed within a single tube that contains a pool of various PCR primers specific for sequences of the genome that flank each individual deletion. The name gap-PCR is derived from the fact that each pair of primers spans the targeted deletion ("gap") breakpoints, such that a product of the expected size is only produced if the deletion is present (Fig. 24.9); without the deletion, the primers flank a region too large to amplify. PCR products are typically detected by standard agarose gel electrophoresis. A limitation of gap-PCR is that it can only detect targeted deletions. If a novel deletion is present that is outside the region of interest targeted by the primer mix, gap-PCR results will be normal. Thus, if there is a high clinical suspicion of alpha-thalassemia, a normal gap-PCR result may be reflexed to an alternative molecular method.

Case Example 5

As part of a routine annual physical exam, an 18-year-old Caucasian male had a CBC performed, which was remarkable for an increased red cell count and decreased MCV (Table 24.8). Further investigation by HPLC showed normal HbA_2 (2.5%), and inclusion body studies noted three or four "possible" HbH inclusions. Gap-PCR was performed to determine if the abnormal red cell indices were due to an Hb abnormality, and it was determined that the patient's genotype is −−SEA/aa. What does this result mean for the patient?

Case 5 Resolution

Alpha-thalassemia disorders are common, with estimations suggesting approximately 5% of the worldwide population carries an alpha-thalassemia variant. Absent or decreased production of the alpha chain of Hb causes alpha-thalassemia, and the severity can range from asymptomatic to lethal. There are two alpha globin genes on each chromosome 16 for a total of four alpha globin genes. Although the presence of four genes complicates

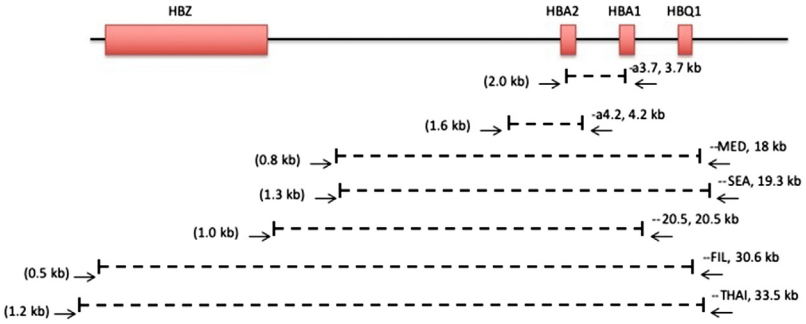


FIGURE 24.9 Gap-PCR schematic for alpha-thalassemia. Dashed lines: portions removed by each deletion. Arrows: primer locations that flank each deletion. The approximate size of each gap-PCR product is shown to the left of each deletion in parentheses.

TABLE 24.8 Laboratory results for Case Example 5.

Hematologic parameter (units)	Result	Reference interval
WBC ($10^3/\mu L$)	8.2	4.3–10.0
RBC ($10^6/\mu L$)	6.2	4.4–5.6
Hb (g/dL)	14.0	13.0–18.0
HCT (%)	42	38–50
MCV (fL)	67.5	81–98
MCH (pg)	23.1	27.3–33.6
MCHC (g/dL)	33.3	32.2–36.5
Platelet count ($10^3/\mu L$)	280	150–400

Complete blood count remarkable for elevated red cell count with decreased mean corpuscular volume, and mean corpuscular hemoglobin.
Hb, Hemoglobin; *HCT*, hematocrit; *MCH*, mean corpuscular hemoglobin; *MCV*, mean corpuscular volume; *RBC*, red blood cell; *WBC*, white cell count.

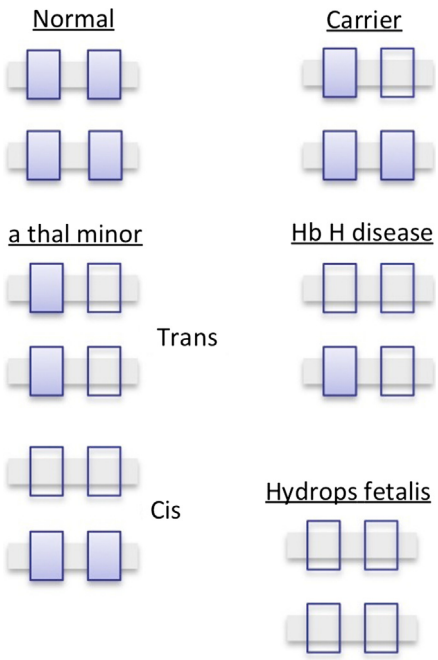


FIGURE 24.10 Four alpha globin genes are depicted. Filled boxes: normal gene. Open boxes: deletion. Note the cis and trans configurations for alpha-thalassemia minor.

genotypic characterization, phenotypes typically correlate with the number of nonfunctional genes. In this case, the patient has an SEA deletion, which involves deletion of two alpha globin genes. Since two alpha globin genes remain intact in this patient, the most significant clinical manifestations for this patient will be mild hypochromic microcytosis.

For genetic counseling purposes, it is important to determine whether the two deletions occur on a single chromosome (−−/aa, cis configuration) or whether one gene is mutated on each chromosome (−a/−a, trans configuration; Fig. 24.10). The cis configuration is at increased risk for producing offspring with HbH disease or hydrops fetalis, because one chromosome entirely lacks functional alpha globin genes. The −−SEA deletion is a cis deletion, and it is historically most prevalent in Southeast Asia where HbH disease and hydrops fetalis are more common. Thus this patient should be counseled that he is at risk for transmitting the −−SEA chromosome to offspring. In the patient's family, there was no history of anemia, and the family was unaware of Southeast Asian ancestry. This case highlights that the incidence of alpha-thalassemia is increasing globally and that molecular determination of the underlying genetic defect has significant implications.

Assays to detect unknown variants

Nonsequencing methods

Historically, a molecular method that has been widely used and does not involve DNA amplification is Southern blot hybridization, which is still utilized in certain contexts. It is particularly helpful when screening for large deletions or rearrangements, and is indispensable for identifying novel deletions. Prior to the advent of PCR, fluorescence in situ hybridization (FISH) analysis was also used to detect large alterations, though FISH probes have much less utility for detecting novel changes. Another alternative method is an array, which can be a single-nucleotide polymorphism (SNP) or comparative genomic hybridization array. Both methods can be used to detect large-scale alterations, such as deletions, and can also detect unknown deletions. For more information on these technologies, please refer to Chapter 13, Nucleic acid analysis in the clinical laboratory.

Two PCR-based methods for detecting unknown mutations exploit altered single-stranded DNA conformation and include denaturing gradient gel electrophoresis (DGGE) and single-strand conformation polymorphism analysis. An advantage of these methods is that they can detect a single-nucleotide change as well as small insertions and deletions. Although these methods are still viable, they require technical expertise and are labor-intensive, which increases overall turnaround time. Thus many laboratories have discontinued these methods in favor of sequencing.

Multiplex ligation-dependent probe amplification

Another PCR-based method that can detect uncommon and novel large rearrangements and copy number

alterations is MLPA. This method relies on multiplex probes that are available as a kit for either the alpha or beta globin gene clusters, or laboratories can design a custom probe mix. An advantage of this method is that the necessary instrumentation, including PCR and capillary electrophoresis, is typically available in clinical molecular labs. Briefly, DNA probes are tiled across a region of interest, hybridized to patient DNA, and then amplified selectively, such that PCR product abundance is proportional to the copy number of the target sequence. Products are separated by capillary electrophoresis, and analysis involves comparing a patient's relative peak abundance to characterized controls. A limitation of this method is that it does not identify specific breakpoints for detected alterations. Since definitive identification of the mutation is a key for genetic counseling, one approach that can be used is to screen with MLPA and follow up positive cases with sequencing.

Chain-terminating dideoxynucleotide sequencing (Sanger sequencing)

The relatively small size of the beta and two alpha globin genes makes them quite amenable to amplification and sequencing. Sequencing is an ideal technique for the detection of single-nucleotide changes as well as small insertions and deletions. An advantage of sequencing is that it can readily detect novel mutations as well as those previously described. A challenge with sequencing is that uncharacterized variants may be identified. The international globin gene server database (HbVar: http://globin.cse.psu.edu) is a valuable resource when novel variants are encountered, and family studies can also help elucidate whether an identified variant is pathogenic.

Case Example 6

A 24-year-old African American female is referred for investigation of hypochromic microcytosis with a normal HbA_2 and elevated HbF (Table 24.9). Initial molecular tests for alpha-thalassemia by gap-PCR are unremarkable, as are subsequent sequence analysis studies of the *HBB*, *HBA1*, and *HBA2* genes. MLPA is used to investigate a possible beta globin gene cluster deletion, and a large deletion (13.4 kb) is identified that encompasses the delta and beta genes. What is the reason for the microcytosis in this case? What values from the Hb screen are particularly noteworthy?

TABLE 24.9 Laboratory results for Case Example 6.

Hematologic parameter (units)	Result	Reference interval
WBC ($10^3/\mu L$)	11.2	4.5–13.5
RBC ($10^6/\mu L$)	4.8	4.0–5.2
Hb (g/dL)	12.6	11.5–15.5
HCT (%)	33	35–45
MCV (fL)	68.5	77–95
MCH (pg)	20.9	25.0–33.0
MCHC (g/dL)	38	32.2–36.5
Platelet count ($10^3/\mu L$)	358	200–450
HbA (%)	92.2	n/a
HbF (%)	5.3	<2
HbA_2 (%)	2.5	2.0–3.5

Complete blood count remarkable for decreased mean corpuscular volume and mean corpuscular hemoglobin. Fetal hemoglobin was elevated by biochemical methods, and HbA_2 was normal.
Hb, Hemoglobin; *HbA*, adult hemoglobin; *HbF*, fetal hemoglobin; *HCT*, hematocrit; *MCH*, mean corpuscular hemoglobin; *MCV*, mean corpuscular volume; *RBC*, red blood cell; *WBC*, white cell count.

Case 6 Resolution

This case illustrates several features characteristic for individuals heterozygous for delta/beta-thalassemia. Specifically, HbF concentrations are increased in the context of a normal HbA_2. The MCH and MCV are decreased to a lesser extent than is typically observed in other beta-thalassemias, which is due to increased gamma chain synthesis. The delta/beta deletion causes an imbalance between the synthesis of alpha and nonalpha chains that manifests as mild microcytosis. Although there is a wide range of observed HbF concentrations in delta/beta-thalassemia carriers, the notable finding is that HbF is modestly elevated and HbA_2 is normal. HbF is also increased in HPFH; however, the increase is generally more significant in HPFH (up to 35%). Another distinction between HPFH and delta/beta-thalassemia is that HPFH is not associated with microcytosis. Furthermore, HbF in HPFH patients is typically distributed equally within RBCs (pancellular), which contrasts the heterogeneous HbF distribution found in delta/beta-thalassemia. These findings point to a diagnosis of delta/beta-thalassemia, but family studies and molecular analysis are highly useful for confirmation. Identification of the underlying mutation is also helpful for additional family studies that may be clinically indicated.

Next-generation sequencing

Massively parallel sequencing or next-generation sequencing (NGS) has been rapidly adopted in clinical laboratories for numerous applications that range from prenatal screening for chromosomal aneuploidies to

treatment selection in oncology. A key advantage of NGS is that it is capable of detecting single-nucleotide changes, insertions and deletions, structural rearrangements, and larger copy number alterations. Many factors from assay design to bioinformatics processing affect what information can be gleaned from a test, but there is no doubt that NGS can be effectively used in clinical practice.

To date, there have been few studies that have utilized NGS for the detection of Hb disorders. However, the potential to consolidate testing for Hb disorders to a single methodology is both powerful and appealing. As described above, molecular detection of the varied types of mutations that underlie Hb disorders requires numerous methods that are labor-intensive and technically demanding. Although NGS is challenging both technically and interpretively, there is much interest in a single method that is capable of detecting the entire complement of molecular alterations.

There has also been significant interest in prenatal screening and diagnosis of thalassemia. This has partially been prompted by the continuous adoption of prenatal aneuploidy screening from cell-free DNA in maternal circulation in the general obstetric population. There have been limited studies of prenatal screening for thalassemia from cell-free DNA, but several have reported success. As NGS continues to be widely implemented in clinical laboratories, it will likely be applied to screening and diagnostic purposes in both prenatal and postnatal patient populations.

References

[1] J.D. Hoyer, S.H. Kroft (Eds.), Color Atlas of Hemoglobin Disorders: A Compendium Based on Proficiency Testing, College of American Pathologists (CAP), Northfield, Illinois, 2003.

[2] B. Giardine, J. Borg, E. Viennas, C. Pavlidis, K. Moradkhani, P. Joly, et al., Updates of the HbVar database of human hemoglobin variants and thalassemia mutations, Nucleic Acids Res. 42 (2014) D1063–D1069.

[3] D.N. Greene, C.P. Vaughn, B.O. Crews, A.M. Agarwal, Advances in detection of hemoglobinopathies, Clin. Chim. Acta 439 (2015) 50–57.

Further reading

A.M. Agarwal, R.H. Nussenzveig, C. Hoke, T.S. Lorey, D.N. Greene, Identification of one or two α-globin gene deletions by isoelectric focusing electrophoresis, Am. J. Clin. Pathol. 140 (3) (2013) 301–305.

C.G. Arsene, P. Kaiser, R. Paleari, A. Henrion, M. Spannagl, A. Mosca, et al., Determination of HbA_2 by quantitative bottom-up proteomics and isotope dilution mass spectrometry, Clin. Chim. Acta 487 (2018) 318–324.

K. Chan, M.S. Wong, T.K. Chan, V. Chan, A thalassaemia array for Southeast Asia, Br. J. Haematol. 124 (2004) 232–239.

S.S. Chong, C.D. Boehm, G.R. Cutting, D.R. Higgs, Simplified multiplex-PCR diagnosis of common southeast asian deletional determinants of alpha-thalassemia, Clin. Chem. 46 (2000) 1692–1695.

Y.A. Daniel, C. Turner, R.M. Haynes, B.J. Hunt, R.N. Dalton, Rapid and specific detection of clinically significant haemoglobinopathies using electrospray mass spectrometry-mass spectrometry, Br. J. Haematol. 130 (2005) 635–643.

A.E. Gallienne, H.M. Dréau, J. McCarthy, A.T. Timbs, J.M. Hampson, A. Schuh, et al., Multiplex ligation-dependent probe amplification identification of 17 different β-Globin gene deletions (including four novel mutations) in the UK Population, Hemoglobin 33 (2009) 406–416.

D.N. Greene, A.L. Pyle, J.S. Chang, C. Hoke, T. Lorey, Comparison of Sebia Capillarys Flex capillary electrophoresis with the BioRad Variant II high pressure liquid chromatography in the evaluation of hemoglobinopathies, Clin. Chim. Acta 413 (2012) 1232–1238.

C.L. Harteveld, A. Voskamp, M. Phylipsen, N. Akkermans, J.T. den Dunnen, S.J. White, et al., Nine unknown rearrangements in 16p13.3 and 11p15.4 causing alpha- and beta-thalassaemia characterised by high resolution multiplex ligation-dependent probe amplification, J. Med. Genet. 42 (2005) 922–931.

J. Old, C.L. Harteveld, J. Traeger-Synodinor, M. Petrou, M. Angastiniotis, R. Galanello (Eds.), "Molecular Diagnosis" in Prevention of Thalassaemias and Other Haemoglobin Disorders, Thalassaemia International Federation, Nicosia, Cyprus, 2012.

R. Paleari, D. Caruso, P. Kaiser, C.G. Arsene, C. Schaeffer-Reiss, A. Van Dorsselaer, et al., Developing a reference system for the IFCC standardization of HbA_2, Clin. Chim. Acta 467 (2017) 21–26.

T. Papasavva, W.F. van Ijcken, C.E. Kockx, M.C. van den Hout, P. Kountouris, L. Kythreotis, et al., Next generation sequencing of SNPs for non-invasive prenatal diagnosis: challenges and feasibility as illustrated by an application to β-thalassaemia, Eur. J. Hum. Genet. 21 (2013) 1403–1410.

V.G. Sankaran, J. Xu, S.H. Orkin, Advances in the understanding of haemoglobin switching, Br. J. Haematol. 149 (2010) 181–194.

H.A. Saroff, Model for the action of hemoglobin, Proc. Natl. Acad. Sci. U.S.A. 67 (1970) 1662–1668.

C. Shooter, H. Rooks, S.L. Thein, B. Clark, Next generation sequencing identifies a novel rearrangement in the HBB cluster permitting to-the-base characterization, Hum. Mutat. 36 (2015) 142–150.

C. Vermeulen, G. Geeven, E. de Wit, M.J.A.M. Verstegen, R.P.M. Jansen, M. van Kranenburg, et al., Sensitive monogenic noninvasive prenatal diagnosis by targeted haplotyping, Am. J. Hum. Genet. 101 (2017) 326–339.

W. Wang, Y. Yuan, H. Zheng, Y. Wang, D. Zeng, Y. Yang, et al., A pilot study of noninvasive prenatal diagnosis of alpha- and beta-thalassemia with target capture sequencing of cell-free fetal DNA in maternal blood, Genet. Test. Mol. Biomarkers 21 (2017) 433–439.

D.J. Weatherall, The genetic control of protein synthesis: the haemoglobin model, J. Clin. Pathol. Suppl. (R. Coll. Pathol) 8 (1974) 1–11.

B.J. Wild, B.N. Green, A.D. Stephens, The potential of electrospray ionization mass spectrometry for the diagnosis of hemoglobin variants found in newborn screening, Blood Cell Mol. Dis. 33 (2004) 308–317.

D. Williamson, The unstable haemoglobins, Blood Rev. 7 (1993) 146–163.

Self-assessment questions

1. An adult with a normal hemoglobin profile would be expected to have the following hemoglobin species detected:
   a. HbA only
   b. HbA and HbA_2 only
   c. HbA and HbF only
   d. HbA, HbF, and HbA_2 only
2. Quantitative biochemical techniques are better than qualitative ones for detecting which of the following disorders?
   a. alpha-thalassemia
   b. beta-thalassemia
   c. hemoglobin Barts
   d. sickle cell disease
3. To distinguish between an HbS carrier and an HbS carrier with concurrent beta-thalassemia, it is important to evaluate which of the following?
   a. the ratio of HbS to HbF
   b. the absolute value of HbS
   c. the ratio of HbA to HbS
   d. the relative percent of HbA_2
4. At the molecular level, the most common cause of alpha-thalassemia is which of the following?
   a. deletion
   b. insertion
   c. point mutation
   d. premature stop codon
5. Which of the following is a feature of beta-thalassemia minor?
   a. decreased HbA_2
   b. hypochromic RBC
   c. increased HbF
   d. microcytic RBC
6. Chain-terminating dideoxynucleotide or Sanger sequencing is best suited for detecting single-nucleotide changes and small insertions and deletions. Which thalassemic disorder is most commonly caused by a molecular defect detectable by Sanger sequencing?
   a. alpha-thalassemia
   b. beta-thalassemia
   c. gamma−delta-thalassemia
   d. hemoglobin H disease

Answers

1. d
2. b
3. c
4. a
5. c
6. b

Chapter 25

The complete blood count and white blood cell differential

Alisha D. Ware
Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Describe the morphology and function of red blood cells.
- Describe the morphology and function of subtypes of white blood cells.
- Describe the morphology and function of platelets.
- Describe laboratory methodologies for hematologic measurements.
- Understand the disorders that commonly cause abnormal complete blood count and/or differential results.
- Recognize common laboratory errors that occur on hematology analyzers.

The cellular components of the peripheral blood

The peripheral blood contains three main cellular components: red blood cells (RBCs, erythrocytes), white blood cells (WBCs, leukocytes), and platelets (thrombocytes).

Red blood cells

RBCs are the most numerous component of the peripheral blood [1,2]. Their primary role is in respiration of the body's tissues. A normal RBC count is $4.5-6.0 \times 10^{12}$/L in adult men and $4.0-5.0 \times 10^{12}$/L in adult nonpregnant women [3]. Normal RBCs are $\sim 7-8$ μm in diameter and are shaped like a biconcave disc, with a central pallor that comprises approximately one-third of the cell's diameter (Fig. 25.1) [2,4]. On a Romanowsky or Wright–Giemsa stain, RBCs appear pink or red [2]. Mature RBCs lack nuclei and contain hemoglobin, a protein composed of iron that transports oxygen and carbon dioxide [1]. The life span of a normal RBC is ~ 120 days in systemic circulation [4].

Reticulocytes are circulating immature, nonnucleated RBCs newly released from the bone marrow [1,2,4]. These cells take 1–2 days to mature in the circulation. Reticulocytes contain ribosomal proteins and RNA, which may be visualized with supravital stains, such as methylene blue [2,5]. On a Romanowsky or Wright–Giemsa stain, reticulocytes appear polychromatic, or blue-gray (Figure 25.1) [2,4]. The presence of reticulocytes and nucleated red blood cells (nRBCs) in the peripheral blood corresponds with the bone marrow's ability to increase RBC production in response to physiologic stress, although immature nRBCs and reticulocytes are not abundant and infrequently found in the circulation of a healthy individual [1,5].

Platelets

Platelets are anuclear fragments of megakaryocyte cytoplasm [2,4]. A normal platelet count is $150-450 \times 10^9$/μL [3,6,7]. Typically, mature platelets are between 1.5 and 3 μm in size. They contain azurophilic granules, which are comprised of substances necessary for coagulation and inflammatory reactions (Fig. 25.1) [4]. Immature platelets are larger in size (4–7 μm) and may be increased in association with increased platelet loss or peripheral destruction. Giant platelets (>10 μm in size) may be seen in hereditary disorders or neoplastic conditions [4].

White blood cells

WBCs carry out the body's immune functions [1,4]. Circulating WBCs in a healthy individual include neutrophils, lymphocytes, monocytes, eosinophils, and basophils. A normal WBC count in healthy adults is $4.5-11.0 \times 10^9$/L [3,8]. Neutrophils, also known as polymorphonuclear neutrophils (PMNs), are the most abundant subtype of circulating WBCs in healthy individuals (40%–75%) [2,3]. Mature PMNs are 10–15 μm in size. They have

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00025-9

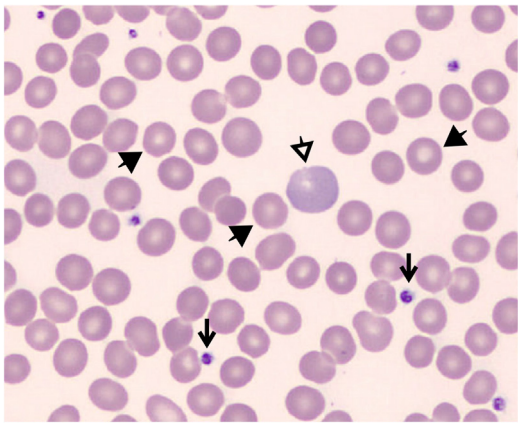


FIGURE 25.1 Wright's stain of a peripheral blood smear. Red blood cells (thick arrows) show central pallor that comprises approximately one-third of the cell's diameter. Reticulocytes (open arrow) may be present in small numbers in normal circulation. Platelets (thin arrows) are typically the smallest circulating cellular component, and they contain numerous azurophilic granules.

granular cytoplasm and a segmented nucleus with two to five lobes connected by a thin nuclear filament (Fig. 25.2A). Band forms, which are less mature neutrophils, contain a single-lobed U-shaped nucleus (Fig. 25.2B) [2,4]. PMNs play a vital role in the body's defense against bacteria and fungi [2,4]. Other immature forms, including metamyelocytes, myelocytes, and promyelocytes, may be seen, circulating in low numbers (<0.5% of circulating leukocytes). An increase in myeloid precursors at various stages in the peripheral blood, termed left shift, may be seen in cases of severe infection, hematopoietic neoplasms, or in other pathologic states [2,4].

Lymphocytes measure 7−15 μm and typically comprise 20%−45% of circulating WBCs [2,4]. Mature lymphocytes have scant pale blue cytoplasm and a round-to-oval nucleus with mature chromatin (Fig. 25.2C) [2−4]. Reactive or large granular lymphocytes have more abundant cytoplasm and fine azurophilic granules. Lymphocytes are involved in defense against viral infections and also play a role in chronic inflammatory conditions [4].

Monocytes are the largest normal circulating cells (12−20 μm) and typically comprise 2%−10% of circulating WBCs. They have gray-blue cytoplasm with an indented or reniform nucleus (Fig. 25.2D). Reactive or neoplastic monocytes may contain azurophilic granules or vacuoles within the cytoplasm [2−4]. Monocytes are important antigen-presenting cells, and are vital in immune modulation and defense against pathogens [4].

Eosinophils comprise approximately 1%−6% of circulating WBCs. They have characteristic brightly eosinophilic (orange-red) cytoplasmic granules and a bilobed nucleus (Fig. 25.2E) [2,4]. Eosinophils are approximately

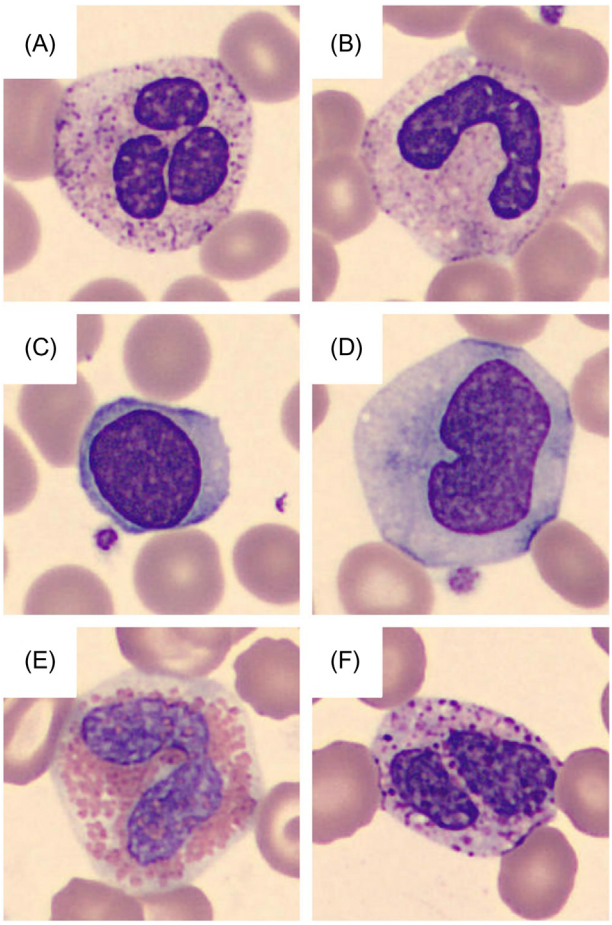


FIGURE 25.2 White blood cell morphology on a Wright's stain of a peripheral blood smear. Segmented neutrophils (A) are typically the most frequent white blood cell type in normal circulation. Band neutrophils (B), which are less mature, are also frequently seen. Lymphocytes (C) are smaller white blood cells with less abundant basophilic cytoplasm. Monocytes (D) have a folded or kidney bean-shaped nucleus and abundant blue-gray cytoplasm. Eosinophils contain numerous orange-red cytoplasmic granules and a bilobed nucleus (E). Basophils (F) also commonly have a bilobed nucleus; however, they contain numerous basophilic granules that may obscure the nucleus.

10−15 μm in size [3,4]. A relative and/or absolute increase in circulating eosinophils is primarily associated with allergic disease, parasitic and fungal infections, and some neoplastic conditions [2,4].

Basophils play a role in allergic disease and infections. In a healthy individual, basophils comprise <1% of circulating WBCs. They are 10−15 μm in size and are characterized by dense purple granules that often obscure the segmented nucleus (Fig. 25.2F) [2,4].

Blasts are immature cells that may be of myeloid or lymphoid lineage, or, rarely, of erythroid or megakaryocytic lineage. Blasts characteristically have a high nuclear-to-cytoplasmic ratio, fine chromatin, and prominent nucleoli (Fig. 25.3). Myeloid blasts may show cytoplasmic granules or Auer rods; however, it is often

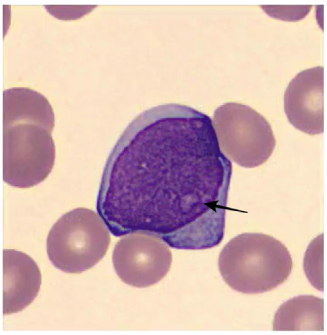


FIGURE 25.3 Blast morphology on a Wright's stain of a peripheral blood smear. Blasts are characterized by a high nuclear-to-cytoplasmic ratio, fine chromatin, and prominent nucleoli (arrow).

difficult to determine blast lineage on morphology alone. Blasts should be exceedingly rare in the circulation, an increase in circulating blasts is concerning for acute leukemia or other bone marrow disorder, and prompt evaluation by flow cytometry is warranted [2,4].

The complete blood count and white blood cell differential

The complete blood count (CBC) is one of the most commonly ordered laboratory tests. A blood sample for a CBC should be collected in an ethylenediaminetetraacetic acid (EDTA) tube. EDTA is an anticoagulant that functions by chelating calcium and other metal ions, thus preventing the enzymatic reactions of the coagulation cascade. Its use allows for preservation of the cellular components of the peripheral blood for laboratory evaluation [9]. Components of the CBC include the RBC and platelet count and indices, and the WBC count with or without the WBC differential. Abnormalities in results will typically generate an alert, termed a flag. Review of a peripheral blood smear may be required to confirm flagged results [1,4,5,10,11]. Table 25.1 shows the reference ranges for the CBC count and WBC differential.

TABLE 25.1 Typical reference ranges for complete blood count and white blood cell differential in healthy adults [2,3,8].

CBC component	Reference range
White blood cell count	$4–11 \times 10^9$/L
Neutrophils	40%–75%
Lymphocytes	20%–45%
Monocytes	2%–10%
Eosinophils	1%–6%
Basophils	<1%
Red blood cell count	$4.5–6 \times 10^{12}$/L (males)
	$4–5 \times 10^{12}$/L (females)
Hemoglobin	14–17.5 g/dL (males)
	12–15 g/dL (females)
Hematocrit	41%–50% (males)
	35%–45% (females)
Mean corpuscular volume	80–96 fL
Mean corpuscular hemoglobin	27–33 pg
Mean corpuscular hemoglobin concentration	33–35 g/dL
Red blood cell distribution width	11%–15%
Platelet count	$150–450 \times 10^9$/L
Mean platelet volume	9–13 fL

CBC, Complete blood count.

Laboratory techniques used in the complete blood count

Modern hematology analyzers use several methods for cell counting, including impedance, conductivity measurements, light scatter, and fluorescence [5]. Impedance, also known as the Coulter principle, involves passing individual blood cells through an analyzer channel. As each cell passes through the analyzer, an electrical pulse is generated. The number of pulses generated correlates with the cell count, while the height of the electrical pulse correlates with cell volume [5,10,12]. Conductivity measurements with high-frequency electromagnetic current provide information on internal structure, including nuclear-to-cytoplasmic ratio, nuclear density, and cytoplasmic granularity [5]. Flow cytometry methods can be used to assess cell count and volume by light scatter characteristics [1,5,10,11]. Fluorescent dyes may be used to label specific cell types, including reticulocytes, platelets, and WBCs, for quantification [1,5,12].

The red blood cell and platelet counts and indices

In a CBC, the RBC count, hemoglobin, mean cell volume (MCV, also known as mean corpuscular volume), hematocrit, mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), and red blood cell distribution width (RDW) are reported, as are the platelet count, mean platelet volume (MPV), and platelet distribution width (PDW).

Red blood cell and platelet counts

One analyzer channel is most commonly used for both RBC and platelet detection, with volume parameters set to differentiate the two cell types. Anucleate cells with a volume of 2–30 femtoliters (fL) are classified as platelets, whereas those with a volume of 40–250 fL are classified as RBCs [5]. This methodology may lead to misclassification of abnormally sized cells. Large platelets may be classified as RBCs, causing a false increase in the RBC count and decrease in the platelet count. Similarly, fragmented RBCs (schistocytes) or microcytic RBCs may be classified as platelets, causing a falsely elevated platelet count and a spuriously low RBC count [1,5]. In such cases, a review of the peripheral blood smear and the patient's clinical history is warranted in order to evaluate the cause of abnormal cellular morphology. Common variations in RBC shape are discussed in Table 25.2, and common causes of microcytosis are described below (see "Abnormal red blood cell count or indices" section).

Fluorescent platelet count

Some automated hematology analyzers offer a fluorescent channel for platelet counts, which improves the accuracy of platelet quantification in thrombocytopenia, when large platelets are present, and in instances in which there are interferences from microcytic RBCs or RBC and/or WBC fragments. The fluorescent platelet channel may be automatically employed by flagged platelet counts, or based on laboratory-specific predetermined parameters [13]. Fluorescent dyes, such as oxazine, thiazole orange, or polymethine, are used to stain the nucleic acids within the cells. The platelets are analyzed via flow cytometry, using laser light to categorize cells as platelets based on size (based on forward light scatter) and fluorescence intensity. This method is also used to determine the immature platelet fraction and differentiate abnormal platelets, RBCs or fragments, and reticulocytes [13–15].

Reticulocyte count

Reticulocytes are commonly measured by fluorescence following staining with a fluorescent dye, such as new methylene blue, oxazine, auramine O, polymethine, and thiazole orange, which binds to residual RNA in the immature RBCs. The cells are then passed through a detector for quantification [1,5].

TABLE 25.2 Common variations in red blood cell shape [2–5].

RBC abnormality	Morphologic characteristics	Associated conditions
Sickle cells	Boat-shaped or sickle-shaped RBCs	Sickle cell disease
Target cells	Targetoid RBCs	Thalassemia, sickle cell disease, hemoglobin C trait or disease, iron deficiency, and liver disease
Ovalocytes	Long axis is 1–2 times length of short axis	Folate and vitamin B_{12} deficiency
Elliptocytes	Long axis is >2 times length of short axis	Hereditary elliptocytosis, thalassemia, hemoglobin S or C trait, cirrhosis, iron deficiency anemia, and megaloblastic or myelophthisic anemia
Stomatocytes	Slit-like central pallor	Hereditary stomatocytosis, acute alcohol ingestion, cirrhosis, obstructive liver disease, and Rh null disease
Echinocytes (burr cells)	Blunt spicules evenly distributed on the RBC membrane	Liver or renal disease, pyruvate kinase deficiency, and artifact
Acanthocytes	Irregular thorny spicules on the RBC membrane	Postsplenectomy, pyruvate kinase deficiency, hemolytic anemia, renal disease, thalassemia, and abetalipoproteinemia
Schistocytes	Fragmented RBC without central pallor	Microangiopathic hemolytic anemia
Spherocytes	RBC without central pallor	Autoimmune hemolytic anemia and hereditary spherocytosis
Bite cells	Bitten apple appearance	Heinz body anemia and G6PD deficiency
Blister cells	Vacuole on the peripheral RBC membrane	Heinz body anemia and G6PD deficiency
Dacryocyte	Teardrop-shaped RBCs	Bone marrow infiltration or fibrosis, megaloblastic anemia, hemolytic anemia, and hypersplenism

G6PD, Glucose-6-phosphate dehydrogenase; *RBC*, red blood cell.

Hemoglobin

Hemoglobin is measured spectrophotometrically following conversion of the molecule to cyanomethemoglobin [1,5,16]. Following RBC lysis, potassium ferricyanide oxidizes the ferrous ion of hemoglobin to the ferric ion of methemoglobin. Methemoglobin is then converted to cyanomethemoglobin following the addition of potassium cyanide. Absorbance at 540 nm is measured, and the intensity of this signal corresponds to hemoglobin concentration [5]. Hemoglobin is reported in grams per deciliter (g/dL). A normal hemoglobin measurement is 14−17.5 g/dL for adult males and 12−15 g/dL for adult females [3].

Mean corpuscular volume and mean platelet volume

The RBC MCV is a measure of the average volume of the RBCs in a given blood sample. The MCV is typically calculated from the RBC histogram, which is plotted by the hematology analyzer (Fig. 25.4A). The *x*-axis of the histogram corresponds to the RBC volume, measured in fL, and the *y*-axis corresponds to the percentage of RBCs at each specific volume. The volume that corresponds to the peak of the curve equals the MCV [1,5]. Like the MCV, the MPV measures the average volume of the platelets. Determination of MPV is analogous to the measurement of MCV (Fig. 25.4B) [1,5].

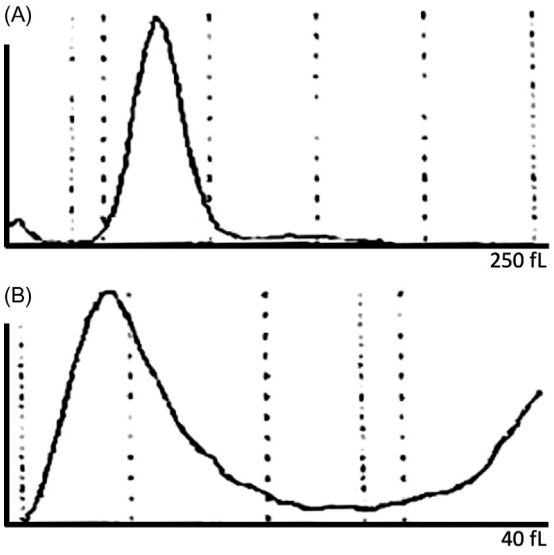


FIGURE 25.4 Red blood cell (A) and platelet (B) histograms produced by a hematology analyzer. The *x*-axis corresponds to cell volume in fL, and the *y*-axis corresponds to the percentage of cells at each specific volume. The mean cell volume of a red blood cell is defined by the volume that corresponds to the peak of the red blood cell histogram curve, and the mean platelet volume equals the volume that corresponds to the peak of the platelet histogram curve.

Hematocrit

The hematocrit is the percentage of packed RBCs in the whole blood [17]. The hematocrit has the implied units of L/L, although it is usually reported as a percent. Prior to the use of modern analyzers, the hematocrit was measured by centrifuging a tube of blood, and measuring the proportion of RBCs, which would settle to the bottom of the tube, to the total volume of blood. On modern hematology analyzers, the hematocrit is a calculated value [5,18,19]. A simple equation to calculate the hematocrit percent is to multiply the RBC count (in cells/L) by the MCV (in fL).

$$\text{Hematocrit} = \text{RBC Count}\left(\text{in RBCs/L}\right) \times \text{MCV(in fL)}$$

Sample calculation for an RBC count of 4.68 × 10^{12}/L and an MCV of 84.8 fL:

$$\text{Hematocrit} = 4.68 \times 10^{12}/\text{L} \times 84.8 \times 10^{-15}\text{L} = 0.397 \text{ L/L(or 39.7\%)}.$$

Mean corpuscular hemoglobin and mean corpuscular hemoglobin concentration

The MCH and MCHC are also calculated on hematology analyzers. MCH, reported in picograms (pg), is a measure of the average amount of hemoglobin in the RBCs [5,18,19]. It can be calculated by dividing the hemoglobin (in g/L) by the RBC count. MCHC measures the average concentration of hemoglobin in the RBCs, and is calculated by dividing the hemoglobin by the hematocrit. Like hemoglobin, the MCHC is reported in g/dL.

$$\text{MCH(pg)} = \text{Hb} \div \text{RBC count}$$

Sample calculation for an RBC count of 4.68 × 10^{12}/L and a hemoglobin of 133 g/L:

$$\text{MCH(pg)} = 133\text{g/L} \div 4.68 \times 10^{12}/\text{L} = 28.4 \text{ pg}$$

$$\text{MCHC}\left(\text{g/dL}\right) = \text{hemoglobin} \div \text{hematocrit}.$$

Sample calculation for a hemoglobin of 13.3 g/dL and a hematocrit of 0.397:

$$\text{MCHC}\left(\text{g/dL}\right) = 133\text{g/dL} \div 0.397 = 33.5\text{g/dL}.$$

Red blood cell and platelet distribution widths

The RDW and PDW are measures of the degree of variation in RBC or platelet size, respectively, in a sample. The RDW and PDW are essentially coefficients of variation of the RBC or platelet volume, and they are expressed as a percentage.

The white blood cell count and differential

Modern hematology analyzers most often report either a three- or five-part WBC differential. Three-part analyzers often use impedance to categorize WBCs as granulocytes, "MID" or intermediate-sized cells (monocytes, eosinophils, and basophils), or lymphocytes; categorization is based on cell size [10,11]. Features such as radiofrequency conductivity may be utilized to provide additional information about WBCs, such as internal complexity or granularity [10]. Five-part analyzers frequently use light scatter by flow cytometry techniques to quantify WBCs and differentiate them by cell size and complexity. As with impedance, cell counts correspond to the number of times the light source is interrupted. A cell's forward scatter is proportional to cell size, while right-angle side scatter correlates with a cell's internal complexity. Side fluorescence correlates with the amount of nucleic acid found within a cell. Each cell type has a distinctive light scatter and fluorescence pattern, which is measured by the analyzer to produce a WBC differential (Fig. 25.5). A five-part analyzer is able to categorize WBCs as neutrophils, lymphocytes, monocytes, eosinophils, or basophils [1,10,11]. Although a three-part analyzer may provide sufficient information in many clinical settings, most laboratories are now using five-part analyzers that provide a more sensitive WBC measurement and greater accuracy in the WBC differential [10,11]. Basophils may be measured via a separate channel in which RBCs and WBCs are lysed, leaving behind the basophils, which are resistant to cell lysis, and bare nuclei of the lysed WBCs [1,12].

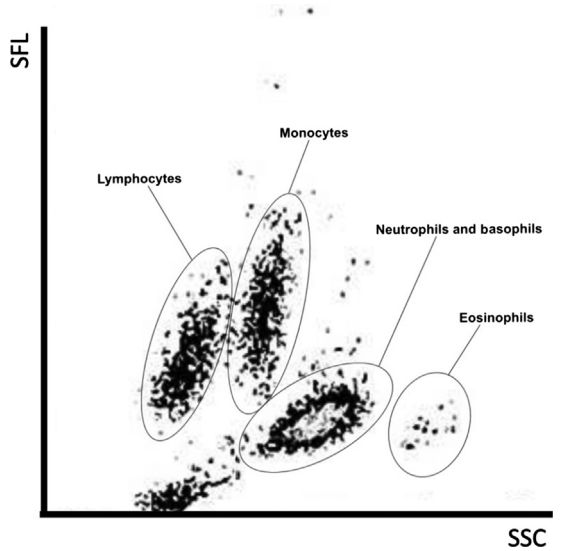


FIGURE 25.5 White blood cell differential plot produced by a five-part white blood cell differential analyzer. Right-angle light scatter, or side scatter, on the *x*-axis corresponds to a cell's internal complexity, while side fluorescence correlates with the amount of nucleic acid within a cell. Each cell type has a distinctive light scatter and fluorescence pattern, as noted on the plot, which is measured by the analyzer to produce a white blood cell differential.

Interpretation of abnormal results for the complete blood count or white blood cell differential

Most modern hematology analyzers will provide an alert, frequently termed a "flag," for abnormal results [5,10,11]. Common reasons for flagged CBC and/or WBC differential results include, but are not limited to, abnormally high or low values, blasts (immature cells), increased immature granulocytes (WBC left shift), atypical lymphocytes, nRBCs, or platelet clumps [10]. Some of the common causes of abnormal CBC and WBC differential results are discussed below. Evaluation of a peripheral blood smear is an essential component in the diagnostic workup of abnormal CBC results, and will be discussed in further detail in the sections to follow.

Abnormal red blood cell count or indices

Anemia

Anemia is defined by the World Health Organization as a hemoglobin concentration of less than 12 g/dL in nonpregnant women, less than 11 g/dL in pregnant women, and less than 13 g/dL in men [5]. Anemia is typically described as microcytic, normocytic, or macrocytic, and can stem from many etiologies including blood loss, bone marrow disorders, genetic disorders, or increased peripheral destruction of RBCs. Microcytic RBCs are abnormally small in size and have a lower than normal MCV ($<$80 fL) due to insufficient production of hemoglobin. Microcytic anemia may occur in iron or copper deficiency, thalassemias, hemoglobinopathies, or bone marrow disorders [1,4,5,20,21]. Normocytic anemia, in which RBCs are normal in size (80−100 fL), may be observed in cases of acute blood loss, RBC destruction, or RBC underproduction. Anemia of chronic disease and sideroblastic anemias may cause normocytic or microcytic anemia. Macrocytic anemia, which is associated with an abnormally elevated RBC MCV ($>$100 fL), may be due to megaloblastic anemia (vitamin B_{12} or folate deficiency), myelodysplastic syndrome, acute blood loss, liver disease, drugs, or alcoholism [2,5,20,21]. In addition to the RBC count and indices, RBC morphology may also aid in determining the cause of anemia (see "Interpretation of the peripheral blood smear" section below). Treatment of anemia involves addressing the underlying cause of disease.

In iron deficiency, the RBCs are microcytic and hypochromic with a low MCHC. The deficiency may be

caused by inadequate intake or decreased absorption of iron, or due to blood loss (i.e., chronic gastrointestinal bleeding). In more advanced disease, the transferrin saturation and ferritin concentrations are low, which may aid in the diagnosis. On bone marrow biopsy, a Prussian blue iron stain will show depleted iron stores [4,5,20,21]. The RDW may be normal or elevated in iron deficiency anemia [1,4,20,21]. The treatment for iron deficiency anemia involves replenishing iron stores with oral supplements or dietary modification [21].

Anemia of chronic disease is a frequent cause of anemia, particularly in hospitalized or chronically ill patients. It most commonly causes normocytic normochromic anemia; however, progressive cases may result in microcytic hypochromic anemia. Anemia of chronic disease may occur in the setting of any chronic disorder, including inflammatory or autoimmune disorders, infections, and malignancy. In anemia of chronic disease, hepcidin, an important regulator of iron homeostasis, is upregulated, which leads to defective iron transport from the gut and inadequate incorporation of iron into erythroid precursors. Laboratory analysis typically reveals a low serum iron, normal or high ferritin, and normal or increased bone marrow iron stores [4,5,20].

Thalassemias cause microcytic hypochromic anemia due to decreased alpha or beta globin chain synthesis. Alpha-thalassemia is caused by deletions of one or more of the four alpha chain genes on chromosome 16, with disease severity corresponding to the number of deleted loci. Beta-thalassemia is caused by mutations in the beta globin gene on chromosome 11. Delta−beta-thalassemia is caused by deletions of segments of DNA on chromosome 11 that involve both the beta and delta globin genes. Disease severity in thalassemia depends on the genetic abnormality, the number of genes deleted or affected, and the presence of a concurrent hemoglobinopathy (i.e., sickle cell disease). Thus a patient with thalassemia may have microcytosis without anemia or may have severe anemia. A diagnosis of thalassemia requires additional laboratory evaluation, including, but not limited to, high-performance liquid chromatography or hemoglobin electrophoresis. However, CBC findings, including a normal or increased RBC count, low MCV, low to normal MCHC, and a normal or increased RDW, can aid in the diagnosis of thalassemia [4,20].

Sideroblastic anemia may result from one of several hereditary or acquired disorders, all of which impair heme synthesis and lead to an abnormal increase in mitochondrial iron. RBCs in sideroblastic anemia may be normochromic and normocytic, or they may be hypochromic and microcytic [4,5,20]. Congenital forms of sideroblastic anemia include X-linked sideroblastic anemia, caused by a mutation in the δ-aminolevulinic acid (ALA) synthase gene leading to a defect in early heme synthesis; X-linked sideroblastic anemia with ataxia, which is caused by a variant in the gene that encodes the ABCB7 transporter protein that leads to defects in heme synthesis or in iron−sulfur biogenesis; and Pearson marrow-pancreas syndrome, which is characterized by large deletions or duplications in mitochondrial DNA, lactic acidosis, exocrine pancreatic insufficiency, and sideroblastic anemia [5,20]. Clonal bone marrow disorders, such as myelodysplastic syndrome with ring sideroblasts, can lead to sideroblastic anemia [5,20,22]. Sideroblastic anemia may also result from drug or toxin exposure, including but not limited to isoniazid, due to inhibition of pyridoxine metabolism; lead poisoning, which inhibits δ-ALA dehydratase and heme synthase; and chronic alcohol abuse, which is directly toxic to erythroid precursors. Serum iron and ferritin concentrations are usually increased [4,5], and a bone marrow examination will show increased iron stores and ring sideroblasts [4,5,20].

RBC underproduction is usually characterized by normocytic normochromic anemia with a paradoxically low or normal reticulocyte count. Causes include pure red cell aplasia, aplastic anemia or other bone marrow syndrome, or bone marrow infiltration by a pathologic process such as tumor, granulomas, histiocytes in storage pool disorders, or fibrosis. Congenital forms of pure red cell aplasia may cause normocytic normochromic anemia; however, they frequently cause macrocytic anemia [4,5,20]. Acquired pure red cell aplasia typically results in a transient decrease in hemoglobin of ∼1 g/dL, and is most frequently due to Parvovirus B19 infection, which is selectively cytotoxic to erythroid precursors. Other viral infections and some therapeutic drugs (i.e., azathioprine, cephalothin, chloramphenicol, dapsone, diphenylhydantoin, isoniazid, and procainamide) may lead to a similar clinical picture. Patients with a hemolytic disorder, such as RBC enzyme deficiencies, membrane abnormalities, hemoglobinopathies, or malaria infection, may also undergo an aplastic crisis with profound anemia. Rarely, antierythropoietin antibodies may form following treatment with erythropoietin; in this case, the red cell aplasia will persist despite discontinuation of erythropoietin therapy, and immunosuppression may be required. Diseases such as thymoma, hematologic malignancies, and autoimmune disorders may lead to chronic acquired pure red cell aplasia; in this case, red cell aplasia is thought to be due to an underlying autoimmune phenomenon. Chronic renal failure may lead to RBC underproduction due to decreased erythropoietin production by the kidneys [20].

Normocytic normochromic anemia may also be seen in the setting of marked RBC loss, such as that seen following hemorrhage or an episode of hemolytic anemia. A reticulocytosis is typically observed 3−5 days following hemorrhage. Hemolytic anemia may be caused by hemoglobinopathies (i.e., sickle cell disease or hemoglobin C

disease) or unstable hemoglobin, intrinsic RBC defects (i.e., hereditary spherocytosis, hereditary elliptocytosis, or hereditary pyropoikilocytosis), RBC enzyme defects (i.e., glucose-6-phosphate dehydrogenase or pyruvate kinase deficiency), plasma factors (i.e., alloantibodies or autoantibodies or drug–antibody complexes), paroxysmal nocturnal hemoglobinuria, mechanical or thermal RBC damage (i.e., microangiopathic hemolytic anemia or following severe thermal burns), or infection (i.e., malaria, babesiosis, and bartonellosis). Other laboratory findings in hemolytic anemia may include reticulocytosis, a decreased RBC life span, increased erythropoietin, indirect hyperbilirubinemia, low haptoglobin, and jaundice/icterus [5,20,21]. In microangiopathic hemolytic anemia, schistocytes are frequently seen on the peripheral blood smear [5,6].

Macrocytic anemia may be observed in bone marrow disorders that result in ineffective hematopoiesis, such as Diamond–Blackfan syndrome [4,5,20], congenital dyserythropoietic anemia [21], and myelodysplasia, or in megaloblastic anemia (vitamin B_{12} or folate deficiency) [4,20,21]. When macrocytic anemia results from a bone marrow disorder, the reticulocyte count is usually low and the RDW is normal [20]. Megaloblastic anemia is caused by vitamin B_{12} and/or folate deficiency, and may result from inadequate dietary intake or decreased gastrointestinal absorption. In patients with vitamin B_{12} deficiency, homocysteine and methylmalonic acid concentrations are markedly increased, while patients with folate deficiency will show increased homocysteine concentrations [21]. Hypersegmented neutrophils are also often seen in megaloblastic anemia [4]. Macrocytic anemia may also be seen following hemorrhage, hemolysis, or treated anemia due to an increase in reticulocytes; in this case, the RDW and reticulocyte count are elevated. Therapeutic drugs, such as those that inhibit DNA synthesis, anticonvulsants, and oral contraceptives, may also cause macrocytic anemia [4,20].

Polycythemia

An increased hemoglobin, referred to as polycythemia, may be seen in myeloproliferative neoplasms (MPNs), such as polycythemia vera (PV), in patients with chronic respiratory disease, or in patients who smoke [4,5]. The hematocrit and RBC count are also elevated in this condition. Patients may present with sequelae of the increased RBC mass, such as hypertension, thrombosis, and bleeding, which can be fatal. Patients with PV may also show leukocytosis and/or thrombocytosis in the early polycythemic phase, and, as the disease progresses, cytopenias often develop due to bone marrow fibrosis [5,20]. A bone marrow biopsy and molecular or cytogenetic testing (i.e., testing for the JAK2 V617F mutation observed in most cases of PV) may be warranted to evaluate polycythemia [5,20,22,23]. Treatment for patients with PV involves reducing the risk of thromboembolic complications and may include anticoagulant therapy, cytoreductive agents, or therapeutic phlebotomy [22,23].

Increased red cell distribution width

An increase in the RDW corresponds to increased variation in RBC size. This can be observed in patients with iron deficiency anemia, myelodysplastic syndrome, megaloblastic anemia, mechanical or thermal RBC injury, or in patients who have been recently transfused [4,5,20].

Abnormal platelet count or indices

Thrombocytosis

Increased platelets, or thrombocytosis, can be caused by reactive conditions, such as infection, inflammation, neoplasms, or iron deficiency. Thrombocytosis can also be seen in MPNs, other malignancy, following splenectomy, or during recovery of thrombocytopenia [2,4,5,24]. In the setting of infection, thrombocytosis is often accompanied by leukocytosis, and anemia may also occur. With resolution of infection, the platelet count often normalizes [25]. In contrast, thrombocytosis associated with an MPN, such as essential thrombocythemia (ET), is typically extreme and persistent, with platelet counts often exceeding 800×10^9/L [22,25]. When an MPN is suspected as the cause of thrombocytosis, a bone marrow biopsy and/or molecular or cytogenetic testing (i.e., for JAK2, MPL, or CALR mutations that are seen in ET and other MPNs) may be necessary to fully characterize the disorder [20,22,26,27]. Patients with thrombocytosis are at risk for thrombotic, vascular, and bleeding complications, which may be life-threatening [27]. Patients with thrombocytosis may also present with pseudohyperkalemia due to excessive leakage of potassium from the platelets during or after blood collection [28].

Thrombocytopenia

The term thrombocytopenia refers to a platelet count of $< 150 \times 10^9$/L. Conditions such as bone marrow and/or genetic disorders, infections, increased peripheral destruction, increased splenic sequestration, and certain medications are common causes of thrombocytopenia [2,4–7]. The manifestations of thrombocytopenia depend on the severity of the decreased platelet count. Patients may experience easy bruising when platelets are mildly decreased; however, severe thrombocytopenia (platelets $<$ 10,000/μL) greatly increases the risk for spontaneous mucocutaneous bleeding and life-threatening intracranial hemorrhage or gastrointestinal bleeding [7]. Patients may undergo platelet transfusions for severe thrombocytopenia

or for mild-to-moderately decreased platelet counts in the setting of a surgical procedure or trauma [6].

Diseases or malignancies that involve the bone marrow, such as aplastic anemia, myelodysplastic syndrome, or marrow infiltration by a malignant, infectious, or reactive process, as well as some toxins and drugs (i.e., chemotherapeutic agents, antimetabolites, chlorothiazides, estrogens, ethanol, and ionizing radiation), can lead to suppressed or ineffective thrombopoiesis [5,6]. There are many genetic disorders associated with thrombocytopenia, which typically present in pediatric patients (Table 25.3) [4–6]. Viral infections such as influenza, human immunodeficiency virus (HIV), Epstein–Barr virus (EBV), human herpesvirus-6, and hepatitis C have been shown to cause thrombocytopenia via deleterious interactions with megakaryocytes and/or platelets. Bacterial and protozoal infections have also been associated with thrombocytopenia [5,6,29].

Peripheral destruction of platelets may occur due to immune or nonimmune-mediated mechanisms. Idiopathic thrombocytopenic purpura (ITP) is the result of platelet-directed autoantibodies and suppression of megakaryopoiesis [5,6]. ITP may occur in isolation or may occur in conjunction with autoimmune hemolytic anemia (Evan's syndrome), or in patients with systemic lupus erythematosus, chronic lymphocytic leukemia, HIV infection, or following stem cell transplantation [5]. Treatment may include steroids, intravenous immunoglobulin, anti-(Rh)D antibodies, and, in severe or refractory cases, splenectomy. ITP is usually a diagnosis of exclusion, as laboratory tests for platelet antibodies are neither sensitive nor specific [6].

Drug-induced thrombocytopenia caused by drug-dependent platelet antibodies has been associated with medications, such as quinine agents and sulfonamides. Removal of the causative agent will often improve platelet count; however, some patients may experience severe thrombocytopenia ($<$10,000/μL) and may require treatment with steroids, IVIg, and/or platelet transfusions. Specialized laboratories may offer testing for the presence of drug-dependent platelet antibodies; however, these assays typically require specialized technical expertise [6].

Heparin-induced thrombocytopenia (HIT) is seen in up to 5% of patients who receive heparin and results from antibodies against heparin–platelet factor 4 (PF4) immune complexes, which engage the Fc receptors on platelets, leading to platelet activation, thrombocytopenia, and risk of potentially fatal arterial and venous thrombosis. Thrombocytopenia is observed following prolonged heparin exposure ($>$4 days), and the CBC will show a drop in the platelet count to approximately half of baseline [5,6]. Patients with HIT should be switched from heparin to an alternative anticoagulant [6]. Immunoassays are available to detect antibodies to heparin–PF4 complexes and to detect the presence of platelet-activating antibodies [5,6]. A heparin-induced platelet aggregation study may also be performed [5]. The serotonin release assay is a more specific testing modality that detects the ability of patient antibodies in a serum sample to activate platelets; however, the assay requires technical expertise and is often not available in clinical laboratories [5,6]. The 4Ts score takes into account the degree of thrombocytopenia, the timing of thrombocytopenia relative to heparin exposure, thrombosis or other sequelae of HIT, and the likelihood that another process is causing the thrombocytopenia. This scoring system is a frequent tool used clinically to determine a patient's likelihood of having HIT [6,30].

Neonatal alloimmune thrombocytopenia (NAIT) is another immune-mediated cause of thrombocytopenia. NAIT results from maternal alloantibodies directed against fetal platelet antigens that are inherited from the father. The majority of cases are due to maternal antibodies against human platelet antigen 1a (HPA-1a). The resulting thrombocytopenia may be severe ($<$20,000/μL), and bleeding complications may ensue. Treatment involves IVIg and/or steroids. If severe thrombocytopenia persists after birth, newborns may require platelet transfusions; however, intrauterine platelet transfusions are less commonly performed. Incompatibility between maternal and paternal platelets may be demonstrated in the laboratory to aid in the diagnosis of NAIT [6,31–35].

Posttransfusion purpura (PTP) is a rare disorder that is also caused by anti-HPA-1a antibodies and can occur in either male or female patients. PTP occurs when an HPA-1a-negative patient receives a platelet transfusion from an HPA-1a-positive donor. Severe thrombocytopenia (platelet count $<$10,000/μL) is seen 5–14 days following the platelet transfusion, which results from destruction of the transfused platelets as well as the patient's own platelets [6,36]. Platelet antibody testing may be performed in order to detect antibodies in the patient's plasma or serum (indirect methods), or to detect antibodies that are attached to the patient's platelets (direct methods), and HPA genotyping may also be performed to help guide clinical management and to confirm antibody specificity. IVIg is most commonly used for therapy in patients with PTP. It is reportedly unusual for patients with PTP to have a subsequent episode with future platelet transfusions [36].

Nonimmune-mediated causes of peripheral platelet destruction include thrombotic thrombocytopenic purpura (TTP), hemolytic uremic syndrome (HUS), and disseminated intravascular coagulation (DIC), all of which frequently result in microangiopathic hemolytic anemia. Acquired TTP is caused by autoantibodies to ADAMTS-13, the protease involved in cleavage of high-molecular-weight von Willebrand factor multimers. A congenital

TABLE 25.3 Most common congenital platelet disorders associated with thrombocytopenia [4–6,18].

Platelet disorder	Inheritance pattern	Genetic abnormality	Clinical findings
Congenital amegakaryocytic thrombocytopenia	Autosomal recessive	*MPL* gene mutation (deletions, missense, or nonsense)	Severe thrombocytopenia, absence of bone marrow megakaryocytes, elevated TPO levels, early pancytopenia (type 1, total loss of TPO), or transient platelet increases with later bone marrow failure
Thrombocytopenia with absent radius syndrome	Autosomal recessive	Unknown/multifactorial (1q21.1 microdeletion, RBM8A exon junction subunit mutations have been reported)	Thrombocytopenia that improves with age, decreased bone marrow megakaryocytes, bilateral absence of the radii, tetralogy of Fallot, atrial septal defects, and lactose intolerance
Bernard–Soulier syndrome	Autosomal recessive	Mutations in the genes (*GP1BA*, *GP1BB*, and *GP9*) encoding platelet membrane proteins (GP1b, GPIX, and GPV)	Mild-to-severe thrombocytopenia with abnormal bleeding times and abnormal ristocetin-induced platelet aggregation studies, large platelets, and normal or slightly increased bone marrow megakaryocytes
Gray platelet syndrome	Autosomal recessive	*NBEAL2* gene mutation	Mild-to-moderate thrombocytopenia, platelets lack α-granules (appear gray on Wright–Giemsa stain), myelofibrosis, and splenomegaly
MYH9 mutation-related disorders (May–Hegglin anomaly, Sebastian syndrome, Fechtner syndrome, and Epstein syndrome)	Autosomal dominant (variable penetrance)	*MYH9* (nonmuscle myosin heavy chain) gene mutations	Mild-to-moderate thrombocytopenia, large platelets, Döhle-like inclusions in WBCs, hearing loss, and nephritis
Mediterranean macrothrombocytopenia	Autosomal dominant	*GP1BA* gene mutation	Mild thrombocytopenia and large platelets
Familial platelet disorder with predisposition to AML	Autosomal dominant (variable penetrance)	*RUNX1* gene mutation	Mild thrombocytopenia, impaired collagen- and epinephrine-induced platelet aggregation studies, dense granule storage pool deficiency, increased risk of developing MDS/AML, or other hematopoietic malignancy at a young age
Paris-Trousseau syndrome	Autosomal dominant	*FLI1* gene mutation	Moderate-to-severe thrombocytopenia that improves with age, abnormal platelet function, psychomotor retardation, congenital heart defects, abnormal facies, and dysmegakaryopoiesis with abnormal or giant α-granules
Wiskott–Aldrich syndrome (X-linked microthrombocytopenia)	X-linked	*WAS* gene mutation	Thrombocytopenia, small platelets with a decreased half-life due to splenic sequestration, immunodeficiency, eczema, normal of increased bone marrow megakaryocytes, and development of autoimmune disorders or lymphoma
X-linked macrothrombocytopenia	X-linked	*GATA1* gene mutation	Moderate-to-severe thrombocytopenia, impaired platelet function, increased and dysplastic bone marrow megakaryocytes, dyserythropoiesis, and anemia

AML, Acute myeloid leukemia; *MDS*, myelodysplastic syndrome; *TPO*, thrombopoietin; *WBCs*, white blood cells.

form of TTP (Upshaw–Schulman syndrome) is an extremely rare disorder due to congenital deficiency of ADAMTS-13. In TTP, the platelet count is often markedly decreased (<50,000/μL), and ADAMTS-13 activity may be <10% of normal levels. ADAMTS-13 autoantibodies may be detected in the acquired form of the disease. TTP is a severe disease with a high mortality rate. In addition to severe thrombocytopenia and microangiopathic hemolytic anemia, patients also may experience renal insufficiency, neurologic sequelae including stroke or headache, and fever. Emergent therapeutic plasma exchange should be strongly considered in patients with TTP, and additional treatments may include steroids, cytotoxic agents, or rituximab [6,37].

Typical HUS is associated with a similar clinical picture to TTP, and patients typically present with bloody diarrhea, renal failure, microangiopathic hemolytic anemia, and thrombocytopenia following an infection with a Shiga toxin-producing bacterial pathogen, such as *Escherichia coli* O157:H7 [6] or *Shigella dysenteriae* [37]. In contrast to TTP, typical cases of HUS are treated with supportive care [6,37]. Atypical HUS is a form of HUS caused by genetic abnormalities in alternative complement regulation. In this disease, there is episodic complement-mediated vascular damage and coagulation. Patients do not usually experience diarrhea, and Shiga toxin is not present. Eculizumab, a monoclonal antibody directed against the complement component C5, is the therapy of choice [6,37].

DIC is a consumptive coagulopathy that leads to widespread thrombus formation, causing thrombocytopenia and microangiopathic hemolytic anemia [6,38]. Causes include systemic infections such as sepsis, trauma or major surgical procedures, malignancy, severe pancreatitis, hemorrhage, hepatic failure, and severe toxic or immunologic reactions [38]. Organ failure develops from microvascular clots, and clotting factors are consumed, leading to bleeding diatheses. The prothrombin and activated partial thromboplastin times are prolonged in the majority of patients. The fibrinogen concentrations may be low in a proportion of cases, and the platelet count drops below 50,000/μL in up to 15% of cases. Treatment involves supportive care, mitigation of bleeding and clotting diatheses, and correction of the underlying cause of DIC [6,38].

Abnormal white blood cell count or differential

Leukocytosis

Leukocytosis, or an elevated WBC count, may be observed infections, inflammatory conditions, or neoplasms. The specific WBC subtype that is elevated can help guide the differential diagnosis in leukocytosis (see "The cellular components of the peripheral blood" section above). The presence of blasts may indicate leukemia or a leukoerythroblastic reaction and may require further investigation, including, but not limited to, flow cytometric and molecular or cytogenetic studies performed on the peripheral blood and/or bone marrow and further clinical evaluation [2,4,5]. It is also important to note that, like patients with thrombocytosis, patients with leukocytosis may have pseudohyperkalemia due to excessive leakage of potassium from the WBCs during or after blood collection [28].

Neutrophilia (PMN count >7000/μL) is the most common cause of leukocytosis. An increase in the absolute or relative concentration of PMNs may occur in infections (i.e., bacterial infections), reactive conditions (i.e., following a surgical procedure), chronic inflammatory diseases (i.e., autoimmune inflammatory conditions and hepatitis), bone marrow stimulation (i.e., recovery from bone marrow suppression, hemolytic anemia, or immune thrombocytopenia), following the use of certain medications (i.e., steroids, beta agonists, lithium, epinephrine, or colony-stimulating factors), or postsplenectomy. Congenital or inherited conditions such as hereditary idiopathic neutrophilia, Down syndrome, and leukocyte adhesion deficiency are rare causes of neutrophilia [5,8].

Lymphocytosis (lymphocyte count >4500/μL or >40% of total WBC count) may occur in patients with infections (i.e., pertussis, syphilis, and viral infections), hypersensitivity reactions, or lymphoproliferative disorders [5,8]. Monocytosis (monocyte count >880/μL) is associated with infections (i.e., EBV, protozoal, rickettsial, or fungal infections), or autoimmune disease [5,8]. Myeloid neoplasms such as acute myeloid leukemia with monocytic differentiation, chronic myelomonocytic leukemia, and juvenile myelomonocytic leukemia are also associated with peripheral monocytosis [5,22]. Eosinophilia (eosinophil count >500/μL) is seen in allergic reactions, parasitic or helminth infections, drug-induced hypersensitivity reactions, many dermatologic conditions [5,8], and rare hematologic malignancies (i.e., myeloid or lymphoid neoplasms associated with eosinophilia) [5,22]. Basophilia (basophil count >100/μL) is rare and may be associated with allergic or inflammatory conditions [8], some viral infections (i.e., varicella infection), endocrine conditions (i.e., hypothyroidism and ovulation), and in chronic myeloid leukemia [22].

Leukocytosis may also be an indicator of a bone marrow disorder, such as leukemia. In acute leukemia, greater than 20% of circulating WBCs will be blasts, or immature blood cells. Blast lineage in peripheral blood samples is best determined by flow cytometric or cytochemical studies, as it is often difficult to accurately determine lineage based on morphology alone [4,5,22,39]. Anemia and/or thrombocytopenia are often seen in patients with acute

leukemia due to replacement of the bone marrow by neoplastic cells [5,39]. Blasts may also be increased in other myeloid neoplasms, such as myelodysplastic syndrome, MPNs, or myelodysplastic/myeloproliferative overlap neoplasms [5,22,39], in patients who have received granulocyte colony-stimulating factor, or in a leucoerythroblastic reaction (i.e., in association with severe infection) [8,39]. Chronic leukemias may be diagnosed in asymptomatic patients and most often show an increase in more mature WBCs (i.e., increased circulating myeloid cells in chronic myeloid leukemia, or increased circulating mature lymphocytes in chronic lymphocytic leukemia) [5,8,22]. Lymphoproliferative disorders including some lymphomas may also present with circulating neoplastic cells (leukemic presentation) [5,22]. Patients with suspected hematolymphoid malignancies should undergo urgent evaluation by a hematologist or oncologist to ensure prompt initiation of therapy if indicated [8,22].

Leukopenia

A low WBC count (leukopenia) may be seen in patients with bone marrow disorders, after use of certain medications (i.e., cytotoxic agents, nonsteroidal antiinflammatory drugs, antipsychotics, antiepileptics, antimicrobial agents, sulfonamides, and gold-containing compounds) [5,40], or in the setting of some infections [4,5]. Neutropenia is most frequently medication-related, but may also result from autoimmune or neoplastic diseases. Lymphopenia is observed in association with infections (i.e., viral or bacterial), autoimmune diseases, or advanced HIV infection. Inherited disorders, such as congenital neutropenia syndromes, Wiskott–Aldrich syndrome, and severe combined immunodeficiency disorder, are rare but may also cause leukopenia [5,40]. Treatment for leukopenia involves addressing the underlying cause. Inherited disorders may be treated with gene therapy or hematopoietic stem cell transplantation [40].

Common errors seen with automated cell counters

Although automated cell counters are generally accurate and reliable, some patient or sample variables may lead to errors in CBC results [1,5]. Many of the errors discussed below will lead to flagged results, which will serve as an alert for further laboratory or clinical investigation. A peripheral blood smear should be reviewed for samples with flagged results or with suspected erroneous results and will assist in identifying the cause of the spurious results. Automated instruments should undergo careful calibration, regular quality assurance testing, and routine maintenance to ensure that the instrument is properly functioning and producing accurate and reproducible results [1,3]. The Clinical and Laboratory Standards Institute and the International Committee for Standardization in Hematology have both established guidelines for instrument calibration and regulatory compliance [1,41].

Errors in red blood cell count

RBC counts may be falsely elevated due to increased RBC or WBC fragments or cryoglobulins. Increased large platelets may also cause a spuriously high RBC count, as the enlarged platelets are counted as RBCs due to increased volume [3,5]. RBC counts may be falsely decreased due to microcytosis; in this case, the small RBCs are counted as platelets by the analyzer. Similarly, increased RBC fragments, or schistocytes, may be incorrectly counted as platelets [1,3,5]. RBC clumps can also cause a spuriously low RBC count [3,5].

Errors in the hemoglobin measurement

Hemoglobin measurements may show false elevation in turbid samples, such as those with a high lipemic index, hypergammaglobulinemia, cryoglobulinemia, hyperbilirubinemia, or marked leukocytosis [1,3,5]. Automated instruments will often have the ability to report lipemic or icteric samples. Chronic smokers will also frequently show an increase in circulating carboxyhemoglobin, which is associated with a left shift in the hemoglobin oxygen dissociation curve, and a higher hematocrit with polycythemia as a physiologic compensatory mechanism [3,5]. Increased sulfhemoglobin, seen in patients following treatment with sulfonamides or nitrogen-containing or aromatic amine drugs and in patients with *Clostridium perfringens* bacteremia, may lead to a spuriously low hemoglobin result [3].

Errors in the mean corpuscular volume measurement

Increased RBC clumps, such as those seen in cold agglutinin disease or paraproteinemia, or hyperosmolar states, may lead to a falsely increased MCV [1,3,5]. Leukocytosis or RBC membrane defects may also cause a spurious increase in MCV, while cryoglobulins, abnormally large platelets, and schistocytes may be associated with an erroneously low MCV [3].

Errors in the platelet count

Causes of a spuriously elevated platelet count include increased RBC or WBC fragments or microcytic RBCs, which are incorrectly quantified as platelets [3,5]. In cryoglobulinemia, protein precipitates may be erroneously quantified as platelets, leading to a false increase in the platelet count.

Pseudothrombocytopenia can be seen in patients with increased large platelets, which are incorrectly identified as RBCs, or in platelet clumping [1,5]. The latter may be induced by EDTA due to platelet-specific antibodies that interact with platelets in the presence of EDTA. This finding is of no clinical significance, and it can be ruled out by review of the peripheral blood smear for platelet satellitism of WBCs [4,5,42] or by repeat testing on a blood sample that is collected in a citrate- or heparin-containing blood tube [5,42].

Errors in the white blood cell count

The WBC count may show spurious elevation in the setting of increased nRBCs, which can be incorrectly quantified as lymphocytes, or in the presence of increased and/or enlarged platelets [1,5]. Cryoglobulinemia or cryofibrinogens may also cause a false elevation in the WBC count when protein aggregates are misidentified as WBCs [1,3,5].

Spurious decreases in the WBC count are most often due to WBC agglutination caused by EDTA, cold agglutinins, or increased circulating immunoglobulins [1,3,5]. Increased WBC fragility resulting in smudge cells, such as the phenomenon seen in chronic lymphocytic leukemia, might lead to a false decrease in the WBC count [3].

Interpretation of the peripheral blood smear

An abnormal CBC or WBC differential will often warrant examination of the corresponding peripheral blood smear to confirm results and/or assess for various disease states. In many labs, any flagged CBC results will trigger peripheral blood smear examination, which may also be useful in excluding spurious results. A peripheral blood smear may also be indicated in instances of suspected malignancy, increased blood counts, and unexplained cytopenias, or to assess for a variety of other disease states [2,5].

Many analyzers are equipped with automated slide makers and stainers that will produce blood smear slides directly from a CBC sample tube [1]. Automated digital image analysis modules that employ convolutional neural networks have the ability to classify blood cells, and are often used to screen blood smears [10,43]. A manual differential is also commonly performed in many laboratory settings.

The peripheral blood smear is evaluated for variations in RBC, WBC, or platelet morphology. Examination of the blood smear can also help identify microorganisms and circulating nonhematopoietic cells. The slide should first be examined at low power to assess for overall cellularity, to check for cellular clumping, to find an appropriate area for morphology assessment, and to evaluate for microfilarial infections [5]. Subsequently, each cellular component should be evaluated at high power for morphologic abnormalities.

Red blood cell morphology

Normal RBCs are approximately the size of a lymphocyte nucleus [2,4,5]. RBCs should be evaluated for abnormal size (i.e., macrocytosis or microcytosis), as well as for morphologic abnormalities including variations in size or shape, and for the presence of intracytoplasmic inclusions.

RBCs may show numerous variations in shape including sickle cells, target cells, ovalocytes, elliptocytes, stomatocytes, echinocytes (burr cells), acanthocytes, schistocytes, spherocytes, bite cells, and blister cells (Table 25.2). RBC inclusions such as Howell–Jolly bodies, Pappenheimer bodies, Heinz bodies, basophilic stippling, or crystals, may be seen (Table 25.4) [2,5]. Infectious organisms such as malaria and Babesia have intracellular forms that infect RBCs, leading to characteristic inclusions; therefore a peripheral blood smear examination should include evaluation for such infections in patients with an appropriate clinical history or symptomology [4].

Platelet morphology

Platelets should be assessed for size and granularity. The peripheral blood smear should also be examined for platelet clumps. Large platelets can be seen during increased platelet production or in inherited disorders. Inherited disorders may also lead to production of small or hypogranular platelets [2,4]. Increased platelet clumping and/or satellitism around neutrophils or monocytes may be seen in samples collected in EDTA-containing tubes (see "Errors in the platelet count" section above) [4,5].

White blood cell morphology

In examination of the peripheral smear, WBCs should be assessed for size, granularity, intracytoplasmic inclusions, and the presence of blasts or circulating nonhematopoietic cells. In cases of leukocytosis, the specific subset of WBCs that is increased may give a clue to the underlying disease process (see "Leukocytosis" section above).

Polymorphonuclear neutrophils (PMNs) have fine cytoplasmic granules and typically have two to five nuclear segments connected by a thin membrane. Hyperlobated PMNs can be seen in megaloblastic anemia and myelodysplastic syndromes. Hypolobated, hypogranular, and/or small neutrophils are seen in myelodysplastic syndrome and genetic disorders. Döhle bodies (blue cytoplasmic inclusions), vacuoles, and prominent azurophilic granules may be seen in reactive or neoplastic neutrophils [4,5].

TABLE 25.4 Common red blood cell inclusions [2–5].

RBC inclusion	Morphologic characteristics	Contents	Associated conditions
Howell–Jolly bodies	Single peripheral body	DNA material	Postsplenectomy, megaloblastic anemia, severe hemolysis, and myelophthisic anemia
Pappenheimer bodies	Multiple small cannon ball-like bodies	Iron material within the mitochondria	Myelodysplastic syndrome, sideroblastic anemia, hemolytic anemia, lead poisoning, and sickle cell disease
Heinz bodies	Single peripheral body visualized with supravital dyes	Denatured hemoglobin	G6PD deficiency and oxidative stress
Basophilic stippling	Multiple punctate granular inclusions	RNA material	Lead poisoning, hemolytic anemia, and pyrimidine 5′-nucleotidase deficiency
Crystals	Dense rectangular crystalline structures, may be intracellular or extracellular	Precipitated hemoglobin C	Hemoglobin C or hemoglobin SC disease

G6PD, Glucose-6-phosphate dehydrogenase; *RBC*, red blood cells.

Reactive lymphocytes are often large in size with abundant cytoplasm that may surround nearby RBCs. Cytoplasmic granules can be seen [4,5]. Circulating neoplastic lymphocytes may also be seen. "Smudge cells" are fragile neoplastic lymphocytes associated with lymphoproliferative disorders that are ruptured during peripheral smear preparation. An albumin preparation may be used to preserve lymphocyte morphology in the setting of numerous smudge cells [39]. Like lymphocytes, reactive monocytes can exhibit increased cytoplasmic granules or vacuoles [4].

Blasts should be exceedingly rare in normal circulation, and their presence should raise suspicion for a neoplasm or leucoerythroblastic reaction. A predominance of circulating promyelocytes, which often have abundant granules and/or Auer rods and may have bilobed or "wasp-waisted" nuclei, should raise suspicion for acute promyelocytic leukemia (APL). Patients with APL are at an increased risk of DIC; however, prompt initiation of therapy with all-trans retinoic acid has proved to decrease the risk of DIC in these patients [4,22,38]. Any patient with a suspected hematolymphoid malignancy should promptly be evaluated by hematology/oncology in order to ensure prompt initiation of therapy if indicated. Additional laboratory studies, including flow cytometry and molecular and/or cytogenetic studies, may be warranted to aid in the diagnosis.

References

[1] K.J. Smock, S.L. Perkins, Examination of the blood and bone marrow, in: J.P.A. Greer, A. Daniel, B. Glader, A.F. List, R.T. Means Jr., F. Paraskevas, G.M. Rodgers, et al. (Eds.), Wintrobe's Clinical Hematology, thirteenth ed., Lippincott Williams & Wilkins, Philadelphia, PA, 2014.

[2] A.S. Adewoyin, B. Nwogoh, Peripheral blood film - a review, Ann. Ib. Postgrad. Med. 12 (2) (2014) 71–79.

[3] R.A. McPherson, M.R. Pincus, Henry's Clinical Diagnosis and Management by Laboratory Methods, twenty-third ed., Elsevier, St. Louis, MO, 2017.

[4] E.F. Glassy, Color Atlas of Hematology: An Illustrated Field Guide Based on Proficiency Testing, second ed., College of American Pathologists, Northfield, IL, 2018.

[5] A. Wahed, A. Dasgupta, Hematology and Coagulation: A Comprehensive Review for Board Preparation, Certification, and Clinical Practice, Elsevier, Waltham, MA, 2015, pp. 1–14.

[6] K.J. Smock, S.L. Perkins, Thrombocytopenia: an update, Int. J. Lab. Hematol. 36 (3) (2014) 269–278.

[7] S.S. Sekhon, V. Roy, Thrombocytopenia in adults: a practical approach to evaluation and management, South. Med. J. 99 (5) (2006) 491–498. quiz 9-500, 33.

[8] L.K. Riley, J. Rupert, Evaluation of patients with leukocytosis, Am. Fam. Phys. 92 (11) (2015) 1004–1011.

[9] G. Banfi, G.L. Salvagno, G. Lippi, The role of ethylenediamine tetraacetic acid (EDTA) as in vitro anticoagulant for diagnostic purposes, Clin. Chem. Lab. Med. 45 (5) (2007) 565–576.

[10] D.S. Chabot-Richards, T.I. George, White blood cell counts: reference methodology, Clin. Lab. Med. 35 (1) (2015) 11–24.

[11] Hematology Analyzers: 3-part or 5-part, That is the Question, Boule Diagnostics AB, Spanga, Sweden, 2019.

[12] P. O'Neil, E. Vital, N. Betancourt-Loria, D. Montes, Performance evaluation of the complete blood count and white blood cell differential parameters on the AcT 5diff hematology analyzer, Lab. Hematol. 7 (2001) 116–124.

[13] C. D'Souza, C. Briggs, S.J. Machin, Platelets: the few, the young, and the active, Clin. Lab. Med. 35 (1) (2015) 123–131.

[14] C. Briggs, P. Harrison, S.J. Machin, Continuing developments with the automated platelet count, Int. J. Lab. Hematol. 29 (2) (2007) 77–91.

[15] C. Briggs, S. Kunka, D. Hart, S. Oguni, S.J. Machin, Assessment of an immature platelet fraction (IPF) in peripheral thrombocytopenia, Br. J. Haematol. 126 (1) (2004) 93–99.
[16] H.H. Billett, Hemoglobin and hematocrit, in: H. Walker, W. Hall, J. Hurst (Eds.), Clinical Methods: The History, Physical, and Laboratory Examinations, Butterworths, Boston, MA, 1990.
[17] H. Mondal, D.P. Budh, Hematocrit (HCT). StatPearls [Internet], StatPearls Publishing, NCBI Bookshelf, 2019.
[18] R.A. McPherson, M.R. Pincus, Henry's Clinical Diagnosis and Management by Laboratory Methods, twenty-third ed., Elsevier, St. Louis, MO, 2017.
[19] P.R. Sarma, Red cell indices, in: H. Walker, W. Hall, J. Hurst (Eds.), Clinical Methods: The History, Physical, and Laboratory Examinations, third ed., Butterworths, Boston, MA, 1990.
[20] E.S. Jaffe, D.A. Arber, E. Campo, N.L. Harris, L. Quintanilla-Martinez, Hematopathology, second ed., Elsevier, Philadelphia, PA, 2017.
[21] J.B. Broadway-Duren, H. Klaassen, Anemias, Crit. Care Nurs. Clin. North. Am. 25 (4) (2013) 411–426. v.
[22] S. Swerdlow, E. Campo, N.L. Harris, E.S. Jaffe, S.A. Pileri, H. Stein, et al., WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. Revised, fourth ed., International Agency for Research on Cancer (IARC), Lyon Cedex 08, France, 2017.
[23] A. Parnes, A. Ravi, Polycythemia and thrombocytosis, Prim. Care. 43 (4) (2016) 589–605.
[24] Z. Gatalica, C. Snyder, T. Maney, A. Ghazalpour, D.A. Holterman, N. Xiao, et al., Programmed cell death 1 (PD-1) and its ligand (PD-L1) in common cancers and their correlation with molecular cancer type, Cancer Epidemiol. Biomarkers Prev. 23 (12) (2014) 2965–2970.
[25] S.R. Rose, N.J. Petersen, T.J. Gardner, R.J. Hamill, B.W. Trautner, Etiology of thrombocytosis in a general medicine population: analysis of 801 cases with emphasis on infectious causes, J. Clin. Med. Res. 4 (6) (2012) 415–423.
[26] A.I. Schafer, Thrombocytosis and thrombocythemia, Blood Rev. 15 (4) (2001) 159–166.
[27] A.I. Schafer, Thrombocytosis, JAMA. 314 (11) (2015) 1171–1172.
[28] Q.H. Meng, E.A. Wagar, Pseudohyperkalemia: a new twist on an old phenomenon, Crit. Rev. Clin. Lab. Sci. 52 (2) (2015) 45–55.
[29] C. Flaujac, S. Boukour, E. Cramer-Bordé, Platelets and viruses: an ambivalent relationship, Cell Mol. Life Sci. 67 (4) (2010) 545–556.
[30] A. Cuker, P.A. Gimotty, M.A. Crowther, T.E. Warkentin, Predictive value of the 4Ts scoring system for heparin-induced thrombocytopenia: a systematic review and meta-analysis, Blood. 120 (20) (2012) 4160–4167.
[31] E. Brojer, A. Husebekk, M. Dębska, M. Uhrynowska, K. Guz, A. Orzińska, et al., Fetal/neonatal alloimmune thrombocytopenia: pathogenesis, diagnostics and prevention, Arch. Immunol. Ther. Exp. (Warsz.) 64 (4) (2016) 279–290.
[32] B.R. Curtis, Recent progress in understanding the pathogenesis of fetal and neonatal alloimmune thrombocytopenia, Br. J. Haematol. 171 (5) (2015) 671–682.
[33] S.F. Gunnink, R. Vlug, K. Fijnvandraat, J.G. van der Bom, S.J. Stanworth, E. Lopriore, Neonatal thrombocytopenia: etiology, management and outcome, Expert. Rev. Hematol. 7 (3) (2014) 387–395.
[34] D. Zdravic, I. Yougbare, B. Vadasz, C. Li, A.H. Marshall, P. Chen, et al., Fetal and neonatal alloimmune thrombocytopenia, Semin. Fetal Neonatal Med. 21 (1) (2016) 19–27.
[35] M.M. Kamphuis, N.P. Paridaans, L. Porcelijn, E. Lopriore, D. Oepkes, Incidence and consequences of neonatal alloimmune thrombocytopenia: a systematic review, Pediatrics. 133 (4) (2014) 715–721.
[36] J. Hawkins, R.H. Aster, B.R. Curtis, Post-transfusion purpura: current perspectives, J. Blood Med. 10 (2019) 405–415.
[37] J.J. Shatzel, J.A. Taylor, Syndromes of thrombotic microangiopathy, Med. Clin. North. Am. 101 (2) (2017) 395–415.
[38] S. Gando, M. Levi, C.H. Toh, Disseminated intravascular coagulation, Nat. Rev. Dis. Prim. 2 (2016) 16037.
[39] D.S. Chabot-Richards, T.I. George, Leukocytosis, Int. J. Lab. Hematol. 36 (3) (2014) 279–288.
[40] C. Onuoha, J. Arshad, J. Astle, M. Xu, S. Halene, Novel developments in leukopenia and pancytopenia, Prim. Care. 43 (4) (2016) 559–573.
[41] England J.M., Rowan R.M., Bull B. S., Coulter W. H., Groner W., Jones A.R., Koepke J.A., et al., ICSH recommendations for the analysis of red cell, white cell and platelet size distribution curves. Methods for fitting a single reference distribution and assessing its goodness of fit. International Committee for Standardization in Haematology. ICSH Expert Panel on Cytometry, Clin. Lab. Haematol. 12 (4) (1990) 417–431.
[42] F. Silvestri, L. Virgolini, C. Savignano, F. Zaja, M. Velisig, M. Baccarani, Incidence and diagnosis of EDTA-dependent pseudothrombocytopenia in a consecutive outpatient population referred for isolated thrombocytopenia, Vox Sang. 68 (1) (1995) 35–39.
[43] A. Acevedo, S. Alférez, A. Merino, L. Puigví, J. Rodellar, Recognition of peripheral blood cell images using convolutional neural networks, Comput. Methods Programs Biomed. 180 (2019) 105020.

Self-assessment questions

1. Which compound is measured to determine hemoglobin concentration?
   a. methemoglobin
   b. sulfhemoglobin
   c. cyanomethemoglobin
   d. carboxyhemoglobin
2. All of the following are methods used to determine the complete blood count except:
   a. impedance
   b. mass spectrometry
   c. flow cytometry
   d. conductivity
3. The red blood cell or platelet distribution width is a measure of which of the following?
   a. variation in size
   b. variation in shape
   c. variation in cytoplasmic complexity
   d. circumference
4. In flow cytometry, the forward scatter is proportional to which cellular parameter?
   a. cytoplasmic complexity
   b. volume
   c. autofluorescence
   d. CD45 expression
5. True or False: Iron deficiency anemia may cause macrocytic anemia.
6. The platelet count may be falsely increased due to which of the following?
   a. platelet clumping
   b. cryoglobulinemia
   c. large platelets
   d. low white blood cell count
7. EDTA may cause all of the following except:
   a. platelet clumping
   b. white blood cell clumping
   c. platelet satellitism
   d. red blood cell fragmentation
8. A predominance of circulating immature cells should raise suspicion for:
   a. infection
   b. acute leukemia
   c. chronic lymphoma
   d. anemia
9. Megaloblastic anemia may result in which of the following? (Choose all that apply.)
   a. large red blood cells
   b. large white blood cells
   c. anemia
   d. leukocytosis
10. An abnormal complete blood count or white blood cell differential result may lead to which of the following?
    a. review of a peripheral blood smear
    b. a result flag
    c. a manual differential
    d. all of the above

Answers

1. c
2. b
3. a
4. b
5. False
6. b
7. d
8. b
9. d
10. d

Chapter 26

Hemostasis

Sean Campbell

Department of Pathology, Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, NY, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Describe the progression of hemostasis from initial activation to fibrinolysis.
- Describe the action of platelets in hemostasis.
- Describe the factor proteins and their different roles in hemostasis.
- List ways in which the PT and aPTT relate to the different pathways in secondary hemostasis.
- Discuss the use of secondary testing in coagulation, such as mixing studies and factor assays, to determine common deficiencies and conditions.
- Describe the various common bleeding and thrombotic disorders and their causes.
- List the tests and algorithms used for elucidating bleeding and thrombotic disorders in advanced coagulation laboratories.

The physiology of hemostasis

There are two main ways to conceptualize the hemostasis systems: in vitro and in vivo. Because of the history of coagulation testing, the most common coagulation tests, prothrombin time (PT) and activated partial thromboplastin time (aPTT or PTT), present a somewhat nonphysiologic picture of the hemostasis cascade. As a result, this chapter will review the physiologic mechanisms as well as the laboratory testing associated with the hemostatic system. First, the hemostasis systems will be presented as they exist in the body. In the laboratory testing section, these systems will be described in terms of the aPTT and PT, to best explain how the systems function, as well as how laboratory tests can illuminate disorders within a system. Following this, more esoteric testing and rare disorders will be presented.

In vivo hemostasis can be broken down into three sections: primary, secondary, and tertiary. Primary hemostasis describes the actions of platelets, while secondary hemostasis represents the coagulation cascade (dominated by factor proteins); tertiary hemostasis encompasses the fibrinolytic and anticoagulant systems, which attenuate the coagulation cascade, control runaway coagulation, and dissolve a clot when it is no longer needed.

Primary hemostasis

Primary hemostasis refers to the action of platelets, as well as collagen and von Willebrand factor (vWF), the multimeric proteins responsible for platelet recruitment and activation at the site of vascular injury. When vessel walls are injured, subendothelial collagen and vWF are exposed. Vasoconstriction occurs, allowing platelets more opportunity to attach to the exposed proteins via two main platelet receptor complexes: GP (glycoprotein) Ia/IIa (collagen binding) or GP Ib/IX/V (vWF binding). These two systems act at different locations due to sheer stress, which is the frictional force of blood flow through veins. Shear stress from blood is highest in smaller vessels, where the surface to liquid ratio is highest, and in areas of high flow such as arteries. Collagen binding is weaker and is more critical in areas of low shear stress, such as large veins, while vWF acts in areas of high shear stress such as arteries and capillaries.

Von Willebrand factor

The structure and function of vWF are worth a further look, due to its interesting structural elements as well as its importance in hemophilia testing. vWF is a multimeric protein, with an individual protein size of approximately 250 kDa. The smallest multimer subunit is a dimer; the two proteins are joined via C-terminal disulfide bonds. The dimers are then multimerized through N-terminal disulfide bonds, and vWF multimers range in size from 500 to 20,000 kDa in healthy persons.

vWF is synthesized in both megakaryocytes and endothelial cells. In megakaryocytes, it is packaged into platelet α granules to be released during platelet activation (see below), while in endothelial cells it is

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00026-0

housed in Weibel–Palade bodies. In Weibel–Palade bodies, some vWF is synthesized as "ultra-large von Willebrand factor multimers" (ULVWFs), greater than 20,000 kDa. ULVWF plays an important role in the very early stages of platelet aggregation, allowing for very rapid recruitment and activation of platelets after a vessel injury, especially in areas of high shear stress where platelet recruitment is more difficult due to the high rate of flow. However, the presence of these multimers can lead to spontaneous or overactivation of platelets, and thus must be cleaved shortly after release, once the initial platelet recruitment and activation have taken place. The protein responsible for this cleavage is ADAMTS13 (a disintegrin and metalloproteinase with a thrombospondin Type 1 motif, member 13), also known as a vWF-cleaving protease, and functions in areas of high shear stress. In both platelets and endothelial cells, vWF is conjugated to the FVIII protein, a coagulation cascade protein, which is discussed below in the secondary hemostasis section.

Upon release, the vWF can take two forms depending on the levels of shear stress. In areas of low shear stress, as well as in storage, vWF adopts a compact, globular shape. This form has low activity toward platelets, as most of the GP Ib/IX/V binding sites are internal. In areas of high shear stress, the multimer unwinds to up to 1300-nm long, exposing many GP Ib/IX/V binding sites and increasing its platelet affinity manyfold. This interaction of GP Ib/IX/V and vWF exposes ADAMTS13 binding sites, allowing for the cleavage of ULVWF. This complex function leads to many possible abnormalities through sequence or posttranslational modification changes. These changes will be explored more fully in the pathophysiology section.

Platelet function

Platelets are highly specialized cell fragments produced by megakaryocytes in the bone marrow. Platelets serve three distinct functions in hemostasis: they provide structure and support to clotting, they release additional coagulation factors to support the coagulation cascade, and they release vWF and fibrin to add to the clot matrix formation. Platelets are roughly disc-shaped in circulation. Electron microscopy analysis of platelets has determined the presence of several defined internal structures of note. First, platelets have two main categories of granules: α granules, of which there are approximately 50–80 per platelet and which contain mainly proteins such as fibrinogen, factor V, factor XIII, vWF, and plasminogen; and dense granules, also known as δ granules, of which there are five to seven per platelet and which contain small molecules such as ADP, ATP, Ca^{2+}, and serotonin. In addition to these granules, platelets also contain fine tubules that connect to the platelet surface, termed the open canalicular system (OCS).

Activation, through binding to collagen and vWF, leads to shape changes that form the platelet into a shape more reminiscent of a dendritic cell, with a spherical center and many long pseudopodia. This shape change allows for easier access of the platelet receptors to their ligands, and constriction of the platelet brings the granules in contact with the OCS, allowing for the rapid release of the contents of said granules into circulation. A final step of platelet activation is the inversion of the plasma membrane phospholipids. In the resting state, platelet outer membranes consist nearly entirely of phosphatidylcholine and sphingomyelin, which are not active in the coagulation pathway. Upon platelet activation, phospholipid "scramblases" such as transmembrane protein 16F (also known as anoctamin 6) are activated and move negatively charged phospholipids such as phosphatidylserine from the inner membrane to the outer plasma membrane. These phospholipids serve as anchor points and activators of other coagulation factors, amplifying the hemostatic process at the site of vascular injury.

Platelet receptors

There are over 50 different receptors present on the platelet surface; of these, only a select number have known clinical significance. Of the clinically significant receptors, the G-protein-coupled receptors (GPCRs) GP Ia/IIa and GP Ib/IX/V have already been mentioned, and allow platelets to bind to collagen and vWF, respectively. In addition to these, GP IV and GP VI also participate in collagen binding, though their role is less well-defined than that of GP Ia/IIa. These receptors help anchor the platelet and begin activation and membrane changes through activation of their associated G-proteins (Fig. 26.1).

These receptor proteins activate phospholipase A2 (PLA_2) and phospholipase C (PLC). Both enzymes act on phosphatidylinositol (PI) in the platelet's inner membrane. PLA_2 releases arachidonic acid (AA) from PI. AA, in turn, is converted into thromboxane A_2 (TXA_2) through the actions of cyclooxygenase (COX) and thromboxane synthase. TXA_2 is released by the activated platelets, and can bind to and activate other platelets through the thromboxane/prostanoid receptor isoform alpha, which inhibits adenylate cyclase (AC), decreasing cyclic adenosine monophosphate (cAMP) and allowing cytoplasmic Ca^{2+} concentrations to rise, as well as directly stimulating calcium release from cell stores through activation of PLC. TXA_2 also spontaneously converts to TXB_2, which is stable and inactive.

Simultaneously, activated PLC forms inositol-1,4,5-triphosphate and diacylglycerol from PI, and these

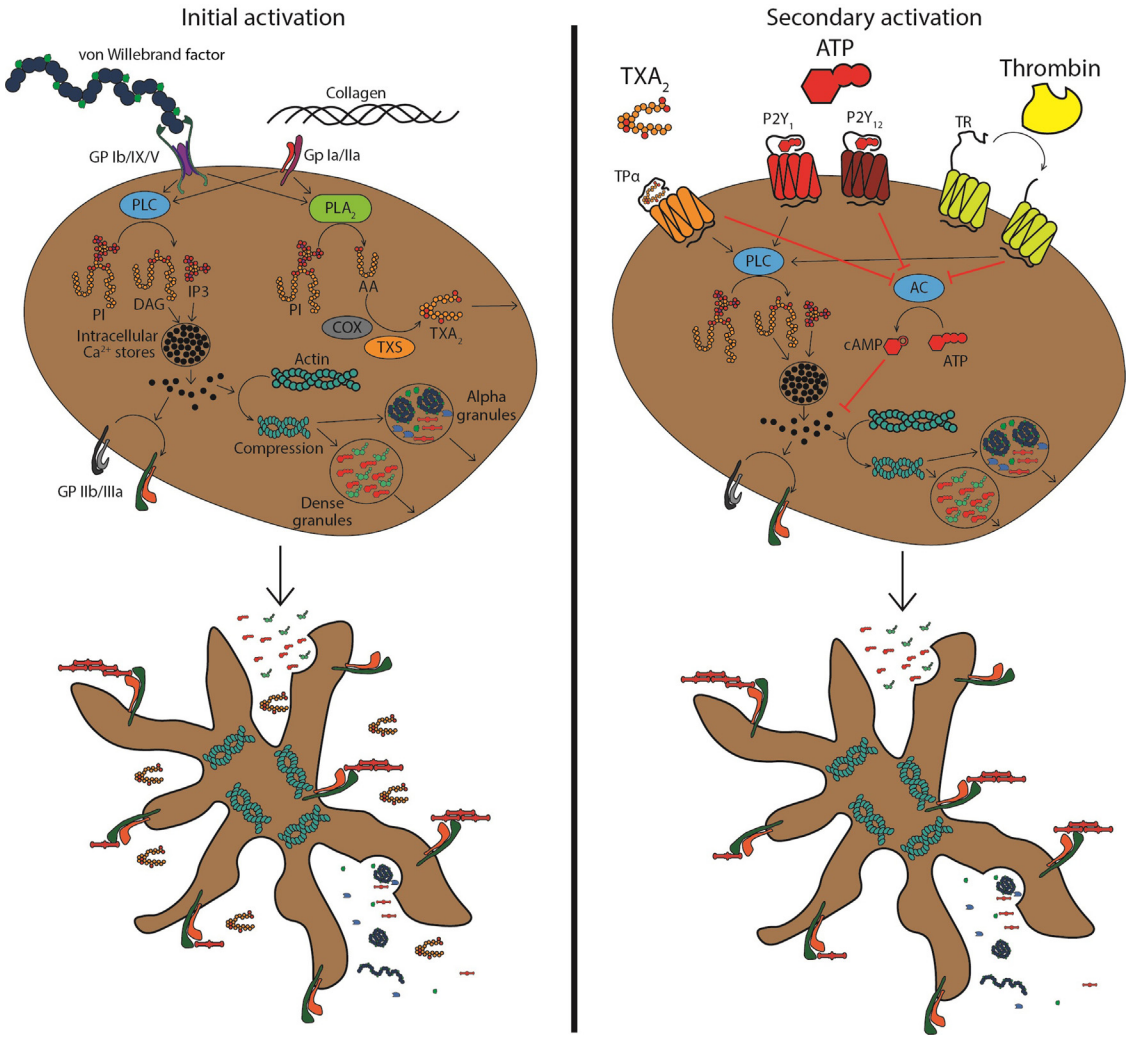


FIGURE 26.1 Pictorial representation of platelet activation. *Left*: During vascular injury, platelets are initially activated through interaction with von Willebrand factor multimers and collagen. These bind to GP Ib/IX/V or Gp Ia/IIa, respectively, which in turn activate phospholipase C and phospholipase A2. Phospholipase C converts phosphatidylinositol into diacylglycerol and inositol-1,4,5-triphosphate. These stimulate release of calcium from intracellular stores. Rising intracellular calcium stimulates actin compression and rearrangement, causing release of granules and membrane changes to change the platelet into a more dendritic shape. It also activates Gp IIa/IIIa, which binds to fibrinogen and anchors the platelet to the growing clot. Phospholipase A2 converts phosphatidylinositol into arachidonic acid, which is then converted to thromboxane A2 through the successive actions of cyclooxygenase and thromboxane synthase (thromboxane A synthase) enzymes. Thromboxane A2 is then released from the platelets. Full activation results in the platelet radically altering shape, releasing granule contents, and binding to fibrinogen. *Right*: After initial activation, release of thromboxane A2 and ADP from activated platelets, as well as active thrombin created from the coagulation cascade, causes a secondary activation of platelets. The three act through associated receptors, thromboxane/prostanoid receptor isoform alpha, $P2Y_1$ and $P2Y_{12}$, and thrombin receptor, to activate phospholipase C and inhibit adenylate cyclase. Adenylate cyclase forms cyclic adenosine monophosphate from ATP, and cyclic adenosine monophosphate suppresses intracellular calcium. Inhibition of adenylate cyclase allows calcium levels to rise, supporting platelet activation.

products directly release intracellular stores of Ca^{2+} and cause actin contraction, leading to pseudopod extension. Ca^{2+} release also results in a conformational change in the GP IIb/IIIa receptor (also known as integrin $\alpha_{IIb}\beta_3$), allowing the platelets to tightly bind to fibrinogen and thereby facilitating platelet-to-platelet interaction within the fibrin clot (Fig. 26.1).

The contraction of actin and pseudopod extension release granules into the OCS, allowing for the activation of additional nearby platelets and circulating coagulation factors. vWF and fibrinogen released from the dense granules help anchor the forming platelet plug, as well as recruit additional platelets to the site of injury. The release of factor V, factor XIII, and Ca^{2+} support the activation of secondary hemostasis. ADP in the α granules activates other platelets through the $P2Y_1$ and $P2Y_{12}$ GPCRs. These, in turn, activate PLC, activating the nearby platelets in a similar way to those activated through GP Ia/IIa, and inhibit adenylate cyclase to support the increase of intracellular calcium.

The thrombin receptor (TR) is the last of these receptors that will be covered in detail. Unlike the GP proteins mentioned previously, the TR is a protease-activated receptor. Cleavage of the receptor N-terminus by thrombin allows for a conformational change to the protein structure, which is then transmitted through the platelet membrane to activate the receptor-associated proteins G_i and G_q, which also stimulate PLC and inhibit AC.

These secondary forms of platelet activation (TXA_2, ADP, and thrombin) lead to further activation and secretion in the platelets, and have the effect of constricting the "core" of the platelet plug into a denser form.

Secondary hemostasis

Secondary hemostasis, also known as the coagulation factor pathway, progresses concurrently to platelet activation, and both contribute to hemostasis in synchrony. Platelets carry many coagulation factors such as factors V and XIII in their granules, and their release helps activate and sustain the signal cascade. In the coagulation factor pathway, activation of prothrombin to thrombin by the coagulation cascade helps to activate platelets further and to form the fibrin matrix that supports the platelet plug.

The factor proteins that control secondary hemostasis are a set of 10 serum proteins typically labeled with Roman numerals: FI (1), FII (2), FV (5), FVII (7), FVIII (8), FIX (9), FX (10), FXI (11), FXII (12), and FXIII (13) (Fig. 26.2). While FI and FII are much more commonly referred to as fibrinogen and prothrombin, respectively, the other factors are exclusively referred to by their factor numbers. An "a" is appended to the proteins to denote their activated form, such as FXa. All factor proteins are protease-activated proteins, though a small fraction (1%–2%) of FVII exists as its active form, FVIIa, in circulation under normal conditions.

There are two essential subgroups of factor proteins: liver-dependent factors and vitamin K-dependent factors. Liver-dependent factors are produced in the liver and thus may be affected by conditions such as cirrhosis and liver insufficiency. These are FV, FVII, FIX, FX, FXI, FXIII, fibrinogen, and prothrombin. Vitamin K-dependent factors

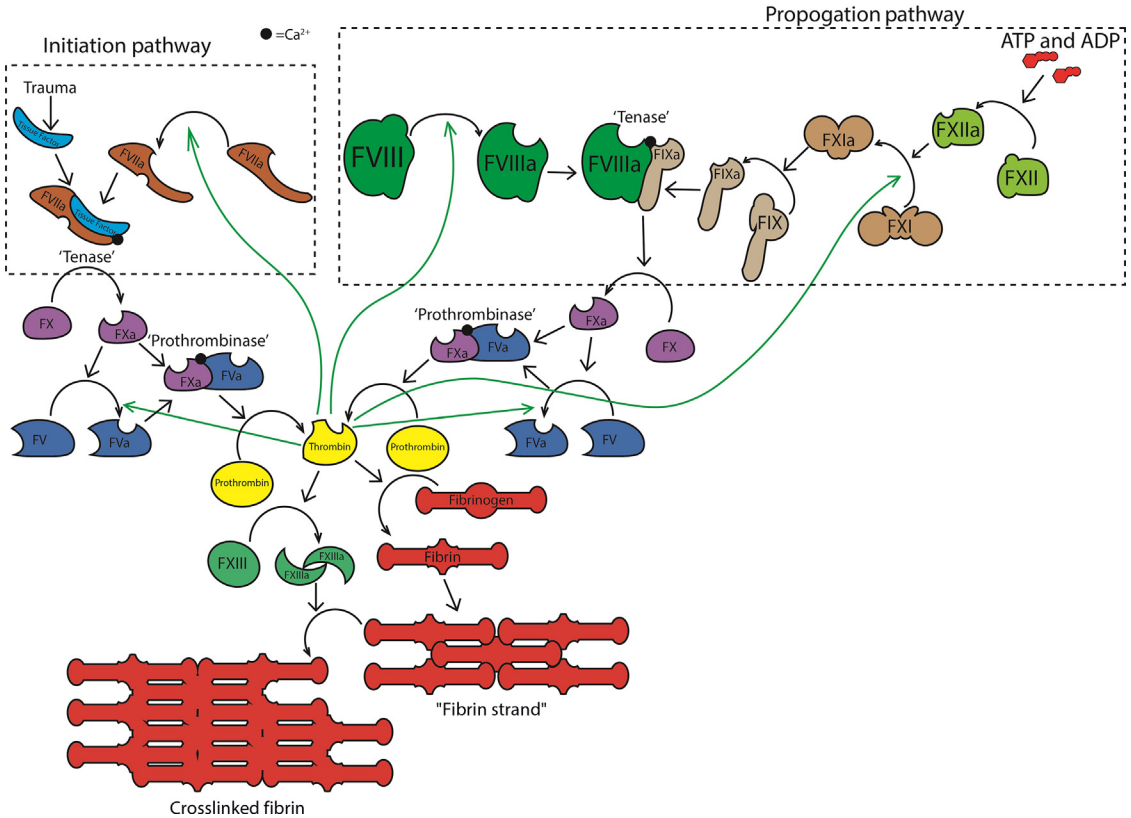


FIGURE 26.2 Pictorial representation of the coagulation factor pathways (secondary hemostasis). The factor pathways are typically initiated by trauma in the initiation pathway (*dashed box, left*). Tissue factor combines with FVIIa to form a tenase complex, which activates FX. The propagation pathway (*dashed box, right*) is initiated on the surface of platelets by the release of polyphosphates such as ADP and ATP. These activate FXII, which activates FXI, which in turn activates FIX. FIXa combines with FVIIIa activated and released from vWF by thrombin to form a different tenase complex to activate FX. FXa generated from either pathway then activates FV and combines with FVa to form prothrombinase, which cleaves prothrombin to thrombin. Thrombin then cleaves fibrinogen to fibrin and FXIII to FXIIIa. Fibrin combines noncovalently with other fibrin molecules to form fibrin strands. These are then covalently bound to form durable cross-linked fibrin by FXIIIa.

all include a *Gla* domain, a conserved protein domain defined by the presence of gamma-carboxyglutamic acid residues. These residues are formed through vitamin K-dependent posttranslational modification of the glutamate residues in a conserved Glu-Xaa-Xaa-Xaa-Glu-Xaa-Cys motif [1]. As the modification is dependent on vitamin K and required for function, these proteins can be affected by malnutrition as well as vitamin K antagonists such as warfarin and superwarfarins like difenacoum (rat poison). Those are FVII, FIX, FX, and prothrombin.

The typical in vivo coagulation cascade starts with vascular injury and the release of a protein termed "tissue factor" (TF). TF is expressed constitutively by many cell types such as fibroblasts, epithelial cells, astroglia, and smooth muscle cells. Importantly, it is not expressed on endothelial cells in the lining of blood vessels; however, it is present in the extracellular matrix around the vessels. The coagulation cascade starts with a vascular injury that releases TF from the extracellular matrix and then proceeds through many steps to the final result, cross-linked fibrin (Fig. 26.2):

1. When blood vessels are damaged, serum from the blood mixes with the extracellular matrix, and TF can come into contact with FVIIa in the serum.
2. FVIIa, TF, and Ca^{2+} form a tenase complex. This complex then cleaves and activates FX to FXa.
3. FXa then cleaves and activates FV to FVa and complexes with it and Ca^{2+} to form prothrombinase. This complex is then cleaves prothrombin to its active form, thrombin.
4. Thrombin then cleaves and activates two other proteins: fibrinogen to its active form fibrin and FXIII to its active form FXIIIa.
5. Fibrin molecules can form noncovalent complexes with other fibrin molecules to form fibrin strands.
6. FXIIIa then cross-links these strands to form a covalently linked polymer and solidify the clot.

Cross-linking fibrin is a very important step in the coagulation cascade, as it allows for the strong base for the mature clot, allowing it to form into a more permanent structure. Changes in fibrinogen structure as a result of mutations or loss of fibrinogen in the blood can lead to highly unstable clots, as platelets adhere to collagen and each other but are unable to solidify into a mature clot.

Fibrinogen is a complex protein consisting of six cross-linked chains, two each of three chains known as Aα, Bβ, and γ. The two Aα chains help form the center of the molecule, also known as the E domain, and the Bβ and γ chains form extensions on either side, termed the D domains (Fig. 26.3). Thrombin cleaves the N-terminal residues of the Aα and Bβ chains, located in the E domain. The cleaved products are known as fibrinopeptides A and B, and leave the α and β chains in the fibrin molecule, now termed a "fibrin monomer." These monomers can noncovalently bind to other fibrin monomers, forming a fibrin strand. This structure is frequently what is measured in coagulation analyzers. Fibrin monomers can then be cross-linked across their D domains by FXIIIa. FXIIIa is a transglutaminase that links γ and α chains to form a covalently linked fibrin clot.

The process described above is the typical initiation of the coagulation cascade in vivo. After initiation, however, in order to strengthen the clot and the clotting response, this clotting cascade will activate a second cascade, known as the propagation cascade, which proceeds through the following steps (Fig. 26.2):

1. Thrombin cleaves FV, FVII, FVIII, and FXI to FVa, FVIIa, FVIIIa, and FXIa. Cleavage of FVIIIa also liberates it from vWF.
2. FVIIa can complex with TF and Ca^{2+} to form more tenase complex to generate additional FXa.
3. FXIa cleaves FIX to FIXa.

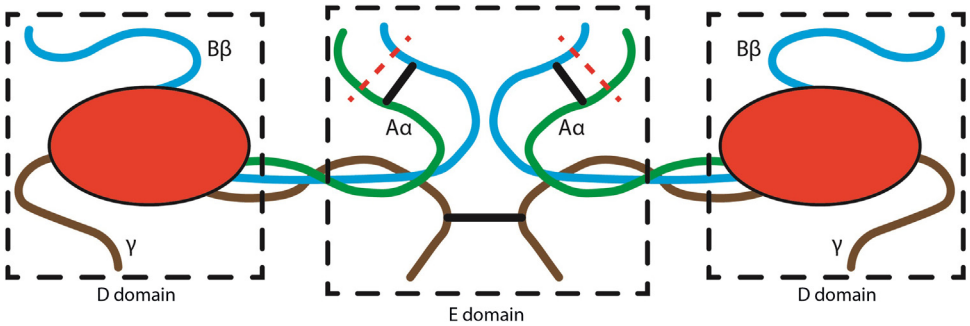


FIGURE 26.3 Schematic of the fibrinogen protein. It is formed through two copies, each of three protein chains: Aα (*green*), Bβ (*blue*), and γ (*brown*). The Bβ and γ chains form the D domain, while the Aα chain coordinates the chains in the center region, termed the "E Domain." The chains are cross-linked across the Aα and Bβ chains, and the two trimers are cross-linked across *y* domains by disulfide bonds to form a single fibrinogen molecule (*black lines*). During coagulation, thrombin cleaves the ends of the Aα and Bβ chains (*red-dashed lines*) to form fibrinopeptides A and B, respectively. This allows the cleaved fibrinogen molecule (now termed a fibrin monomer) to bind to other fibrin molecules noncovalently.

4. FIXa complexes with the FVIIIa activated by thrombin in step 1 and Ca^{2+} to form a tenase complex, which generates additional FXa.
5. The additional FXa generated complexes with FVa cleaved by thrombin to form more prothrombinase.

This additional cascade amplifies the original reaction, allowing for a strong response to the original trauma and expanding the clot to cover the necessary tissue.

Of the factor proteins described earlier, FXII is not mentioned. While initially considered of critical importance to the coagulation cascade, the truth of its necessity was questioned for a period until its physiologic role was discovered. It is now known that, in vivo, FXII is activated by polyphosphates released by platelets, most typically ADP and ATP, as well as phospholipids. FXIIa can then activate FXI, activating coagulation through the propagation stage proteins. As FXIIa cannot activate FVIIIa, the creation of a tenase is very slow without the presence of thrombin, and thus the ability of FXII to start the formation of a clot in isolation is limited. In a forming clot, FXII helps support the propagation cascade. Glass and microsilica particles can activate FXII in a manner similar to polyphosphates, which is important for the aPTT assay discussed later.

Coagulation factors and platelets work in concert, with platelets releasing both factor proteins and additional cofactors, such as Ca^{2+} and the coagulation factors, activating additional platelets. Together, they form the clot that slows bleeding and allows the circulatory system to repair itself (Fig. 26.4). Even small variations in this system can lead to aberrant states of either deficient clotting or unnecessary inactivation.

Tertiary hemostasis

Together, primary and secondary hemostatic processes collaborate to form a thrombus or clot, a mixture of platelets and cross-linked fibrin. However, in the absence of other factors, this process would continue indefinitely. Thus the body also has several mechanisms to both slow and stop the coagulation cascade, so that it does not proceed out of control (anticoagulation), as well as dissolve the clot once it is no longer needed (fibrinolysis). Both of these are part of tertiary hemostasis.

Anticoagulation

The body has several distinct mechanisms of anticoagulation, which function at various points in the coagulation cascade. Many of these also function constitutively, creating a barrier that must be overcome for a thrombus to form, preventing unnecessary coagulation from occurring. Tissue factor pathway inhibitor (TFPI) is a protease inhibitor that, true to its name, inhibits the activity of TF by inhibiting the FVIIa/TF tenase complex and thus inhibits creation of FXa. TFPI is produced constitutively by endothelial cells in the blood vessels and thus acts as a constant control on thrombosis (Fig. 26.5).

Protein C (PC) and protein S (PS) work by inhibiting both prothrombinase and tenase complexes. PC is the functional protein, while PS functions as its cofactor. PS-bound PC is known as activated protein C (APC). APC is a protease that cleaves FVa, and then APC can complex to this cleaved FVa to cleave FVIIIa. This inactivates both the factor proteins and their associated prothrombinases. Both PC and PS are formed by the liver and are dependent on vitamin K for proper function.

Another highly important endogenous anticoagulant is antithrombin (AT), also known as AT III. A serine protease inhibitor, AT works in concert with anionic glycosaminoglycans, the most recognizable of which is heparin. In vivo, AT function is likely potentiated by heparin sulfate, which is a major component of many cell surfaces including endothelial cells. Heparin as a pharmaceutical product is closer to that produced solely by mast cells and is likely not related to coagulation in vivo. The pharmacologic heparin is also known as "unfractionated heparin," in contrast to "low-molecular-weight heparin" (LMWH) which is a more purified form of heparin comprised solely of shorter polypeptides. Unfractionated heparin works by noncovalently joining AT to its targets, thrombin and FXa, which inactivates them. LMWH works by a similar mechanism but is only capable of directly inhibiting FXa. AT is formed in the liver but is not a vitamin K-dependent protein.

Finally, there is protein Z-dependent protease inhibitor (ZPI). ZPI is another serine protease inhibitor, and while it does not require protein Z (PZ) to function, the presence of PZ increases its activity by over 1000-fold. ZPI acts by inhibiting FXa. ZPI and PZ are synthesized in the liver, and PZ is vitamin K-dependent.

Fibrinolysis

Unlike anticoagulation, fibrinolysis is dominated by a single protein: plasmin. Plasmin is a protease-activated serine protease and is the primary driver in the dissolution of a mature clot. Plasmin acts primarily through the degradation of both fibrin and fibrinogen, leading to fibrinogen degradation products (FDPs). D-dimers are a specific type of FDPs that are only found from cleaved, cross-linked fibrin, thus being specific for the action of plasmin on a mature clot. This has implications for testing.

Plasminogen is synthesized in the liver, and its activation to plasmin is started during secondary hemostasis, such that clot degradation is occurring concurrently to the formation. Plasmin is formed through several activation mechanisms, most notably FXIa, FXIIa, and kallikrein.

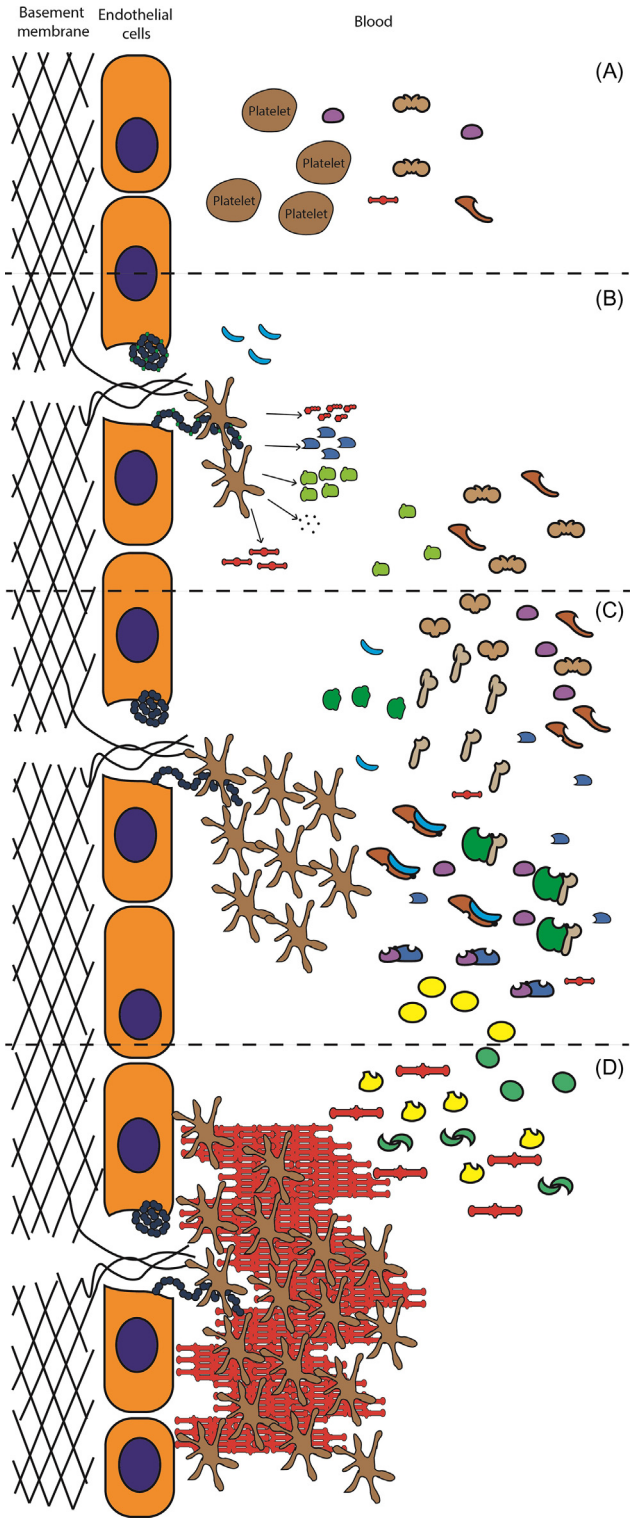


FIGURE 26.4 Pictorial representation of the clotting cascade. (A) No activation has occurred, and factor proteins and platelets both exist in their normal forms. (B) Damage to the endothelial wall has caused the release of collagen and vWF from the basement membrane and endothelial cells, as well as tissue factor. This leads to activation of platelets, which adhere to the site and release factor proteins as well as calcium, ATP, fibrinogen, and other molecules. (C) More platelets have aggregated, and factor proteins are beginning to be activated, leading to the formation of tenase and prothrombinase complexes. (D) Activation of thrombin and subsequently FXIIIa leads to the creation of cross-linked fibrin, which secures the growing clot and leads to a stable protection of the damaged site. Not shown are thrombolytic enzymes, which will break down the clot when it is no longer needed.

(*Continued*)

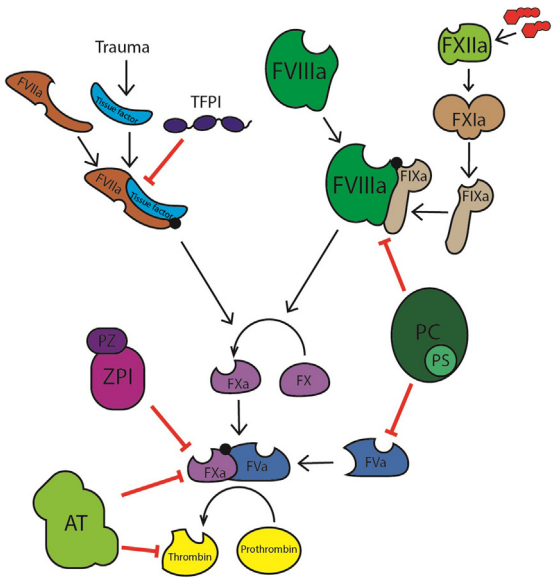


FIGURE 26.5 Simplified diagram of the coagulation system, displaying the action of various anticoagulant molecules (see Fig. 26.1 for full diagram). Tissue factor pathway inhibitor inhibits the tissue factor: FVIIa tenase complex. Protein Z-dependent protease inhibitor inhibits FX (shown in complex with protein Z). Protein C and S inhibit FX as well as FVIII. Antithrombin inhibits factor X as well as its namesake, thrombin.

Kallikrein is a serine protease and is cleaved to its active form by FXIIa. Plasmin can also be formed by the activity of the aptly named tissue plasminogen activator (tPA). tPA is synthesized and released by endothelial cells.

Laboratory testing

The goal of laboratory testing in hemostasis is to diagnose issues in this complex system quickly and accurately. As such, there is a general sequence of testing that occurs for most patients with coagulopathies, as well as most hospitalized patients. This testing typically starts with the most basic coagulation pathway tests, the aPTT (or PTT) and the PT. Both proceed through similar steps, the main difference being the activation process.

For coagulation tests, draw volume is essential. The final concentration of citrate must be accurate, as, unlike other plasma tests, the anticoagulation needs to be reversed. In most cases, coagulation samples are citrated plasma from whole blood collected into tubes with 3.2% (109 mM) liquid sodium citrate. Citrate removes the

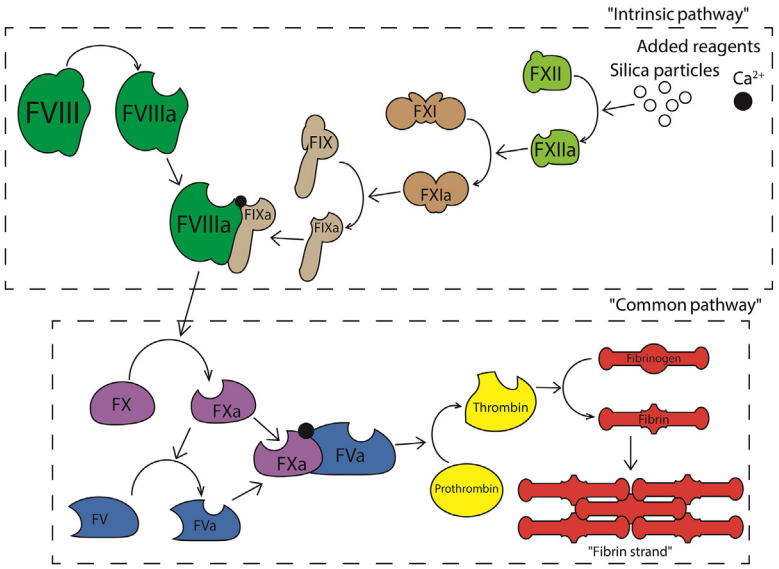


FIGURE 26.6 Pictorial representation of the activated partial thromboplastin time pathway. Silica particles (or another activated partial thromboplastin time activator such as kaolin) and calcium are added to an anticoagulated sample. This activates endogenous FXII, which activates FXI, which in turn activates FIX. FIXa combines with FVIIa and calcium to form a tenase complex. This is known as the intrinsic pathway. This tenase complex then activates FX to FXa, which in turn cleaves and activates FV to FVa. FXa then combines with FVa and calcium to form a prothrombinase complex that cleaves and activates prothrombin to thrombin. Thrombin cleaves fibrinogen to fibrin, which noncovalently binds other fibrin molecules to form fibrin strands. This is known as the common pathway, as it is also activated in the prothrombin time assay. The formation of these fibrin strands is the endpoint of the activated partial thromboplastin time assay.

calcium ions required for coagulation factors to bind to phospholipid membranes, thereby inhibiting coagulation. The final ratio of anticoagulant to blood should be 1:9, which for most blood collection tubes is approximately 90% full. Given a typical hematocrit of 40%, this results in a final citrate concentration of approximately 17 mM. Every lab performing coagulation testing must define specific limits of specimen acceptability for fill volume and hematocrit in order to ensure that this concentration is accurate. In cases of either underfilling or highly aberrant hematocrit, the final sodium citrate concentration will be incorrect and the reversal will not be successful, resulting in falsely long PT and aPTT results, mimicking a coagulopathy.

Another important preanalytical factor in coagulation testing is centrifugation. Serum coagulation testing is meant to measure the ability of the serum proteins to initiate coagulation, not the platelets that carry large stores of additional factor proteins and procoagulation molecules such as calcium. As a result, centrifugation of citrated samples must be sufficient to generate platelet-poor plasma (PPP), which is a plasma with a platelet count of less than 10×10^9 platelets/L, though it is good practice to strive for less than 5×10^9 platelets/L. Platelets will also be activated by the addition of calcium and prothrombotic factors, and their activation and release of additional coagulation factors will result in falsely low PT and aPTT results, mimicking a thrombotic condition.

Activated partial thromboplastin time

The aPTT (or PTT) measures what is sometimes called the "intrinsic pathway" in addition to the "common pathway," which is a result of the method of collection of serum samples. Historically, samples for aPTT testing were collected in glass tubes. The positive charges of the glass surface caused activation of FXII, and thus the aPTT was believed to be the "intrinsic" method of coagulation, since it was observed spontaneously. In current instrumentation, additives such as silica particles are used to activate this pathway (Fig. 26.6).

Most coagulation tests proceed through a series of common steps. First, the sample is warmed to 37 degrees to promote coagulation. Then, a procoagulant is added, which in the case of the aPTT is one of the mentioned additives, such as silica particles or the protein kaolin, in addition to phospholipids. Calcium is then added in order to overcome the citrate anticoagulant. The sample is incubated and measured for signs of coagulation. Depending on the specific instrumentation used, this may be assessed through either measurement turbidity (optical detection) or viscosity (mechanical detection). This approach measures the formation of fibrin strands (fibrinogen that has been cleaved by thrombin and noncovalently bound) but not cross-linked fibrin; thus FXIII deficiencies are not measurable by this method.

As the aPTT is initiated through the cleavage of FXII, it is most sensitive to deficiencies in the intrinsic pathway proteins. It is also dependent on factors synthesized by the liver, and thus is more sensitive to liver dysfunction. The aPTT is often used to monitor heparin levels for inpatients, and a nomogram can be developed by the lab to correlate blood heparin levels and aPTT results, though a direct Xa activity level or anti-Xa assay (discussed later) is generally considered to be a more precise and accurate assay of heparin activity.

Prothrombin time

The PT is performed in a similar manner as the aPTT, the only difference being the initiation of coagulation. For the

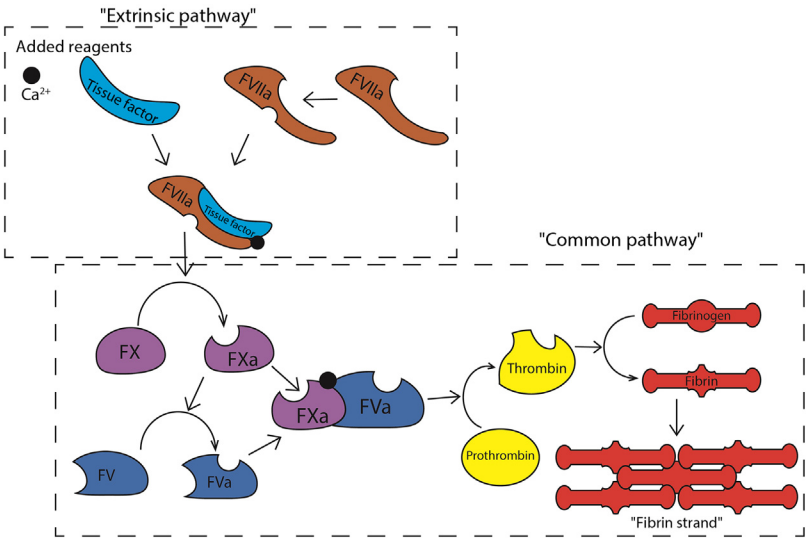


FIGURE 26.7 Pictorial representation of the prothrombin time pathway. Tissue factor and calcium are added to an anticoagulated sample, which combine with endogenous FVIIa to form a tenase complex. This is known as the extrinsic pathway. The prothrombin time then proceeds through the same common pathway as the activated partial thromboplastin time. The formation of fibrin strands is the endpoint of the prothrombin time assay.

PT, coagulation is initiated through the addition of exogenous TF to the plasma, which initiates coagulation through the formation of the TF:FVIIa tenase complex (Fig. 26.7). This TF may be either isolated from animal sources or created recombinantly. Since this approach originally required an additive (unlike the aPTT), this was named the "extrinsic pathway" despite being closer to the in vivo initiation of coagulation. The PT also culminates in the common pathway.

Outside of the initiation phase, this test proceeds in the same manner as the aPTT, though the PT normal ranges are typically much shorter than those of the aPTT as a result of the shorter enzyme cascade required to reach the assay endpoint. In addition, many PT reagents include a heparinase that will neutralize the anticoagulant effect of heparin. As a result, the PT is useful for testing a patient's underlying coagulation state while on heparin but is not useful for heparin monitoring.

International normalized ratio

An important addition to the reporting of the PT is the international normalized ratio (INR). The INR is a calculation that standardizes PT measurement across different reagents. This was created in order to assist in warfarin testing. Warfarin, also known as Coumadin, inhibits vitamin K epoxide reductase, which results in the reduction of previously mentioned vitamin K-dependent factors: FVII, FIX, FX, and prothrombin. Since the extrinsic and common pathways include all of these factors in addition to FV and fibrinogen, the PT is more sensitive to the action of warfarin than is the aPTT. Thus the PT and the associated INR are used as the standard method for warfarin testing.

The INR is a simple calculation, and is typically expressed as $INR = \left(\frac{PT_{specimen}}{PT_{mean}}\right)^{ISI}$. The $PT_{specimen}$ is the result for a given specimen, while the PT_{mean} is the geometric mean PT result for a normal population with the specific assay method, reagent, and lot. The international sensitivity index (ISI) is the numerical value of the assay's sensitivity to warfarin. All manufacturers assign ISI values to their reagents; however, these values must be verified locally in order to ensure their applicability. Verification is conducted through comparison with a reference method and the generation of a log/log plot of the results to determine the slope, which is the ISI. Reagents that are more sensitive to warfarin action will have a lower ISI, typically as low as 0.9, while less sensitive reagents will have higher ISIs, typically as high as 1.5.

Mixing studies

A typical next step in many coagulopathy workups after the discovery of a prolonged PT and/or aPTT is the mixing study. In cases of an isolated factor deficiency, only 50% activity of the factor is required to generate a normal PT or aPTT result. As a result, mixing a patient's sample 1:1 with normal pooled plasma (NPP), defined as pooled plasma with a measured factor activity of 100% in each factor, should normalize a patient's PT and aPTT results even in the absence of a single factor. This is also true for multiple factor deficiencies, such as when a patient is treated with warfarin, though in this case the extent of deficiency is important. Most patients with warfarin in the therapeutic range will correct in a mix. However, this is not the case for inhibitors that will act upon the factors present in NPP and cause the coagulation time of the mixed sample to still be prolonged. Inhibitors may be exogenous, such as heparin or direct oral anticoagulants

(DOACs), or endogenous autoantibodies. In either case, inhibition will cause the mix not to correct back to the normal range for PT or aPTT.

The notable exception to this pattern is in low affinity inhibitors, such as many FVIII inhibitors (seen in acquired hemophilia A; to be discussed later) and some lupus anticoagulants (LAs), which have a unique pattern in mixing studies. Patients with low affinity inhibitors will demonstrate a PT and/or aPTT prolongation (an isolated aPTT prolongation in the case of FVIII and LAs), which corrects upon mixing if the mix is tested right away. However, if the mix is incubated for 2 hours at 37°C, it will again become prolonged. This is because many low affinity inhibitors have slower K_{on} rates, thus requiring incubation in order to see the effects. As a result of this effect, 1- or 2-hour incubations are standard follow-ups for mixing studies that correct in order to confirm the absence of an inhibitor.

Activated clotting time

Activated clotting time [also known as the activated coagulation time (ACT)] is a whole blood clotting assay that is frequently used in point-of-care or near-patient testing. In particular, the ACT is useful in areas where high concentrations of heparin are used, such as cardiac catheterization and extracorporeal membrane oxygenation (ECMO). In these cases, the aPTT is no longer useful, as the concentrations of heparin are such that a clot will not form in PPP. In addition, the possible shorter turnaround time of the ACT as a result of its inclusion in several point-of-care devices makes it useful in operating rooms and cardiac care units. The ACT involves adding a clotting activator, such as celite, kaolin, or glass beads, to a whole blood sample. Clotting is then measured through mechanical (such as measurement of the movement of a metal ball through the serum) or electrochemical (such as measuring electrical resistance of the serum) means. Clotting typically takes between 1 and 3 minutes.

Fibrinogen

Fibrinogen is a key analyte to measure when evaluating coagulation capacity, as it is the functional endpoint of the factor pathways. Fibrinogen may be measured by direct measurement, though the most common method is the Clauss method, a clot-based assay. Most coagulation tests currently performed can be categorized as a "clot-based assay" ("clottable assay") or direct measurement assays (typically called "antigen" assays as many are immunoassays). Fibrinogen is usually performed using a clottable assay, with the endpoint being the formation of fibrin strands.

In the Clauss method, plasma is diluted 10-fold or more, and a large excess of exogenous thrombin is added, as well as calcium, and the time to clot formation is measured. There is an inverse relationship between fibrinogen concentrations and clot time, and a standard curve can be generated through the use of calibrators. Because of the high concentrations of thrombin with respect to plasma, thrombin inhibitors typically do not affect the fibrinogen assay. However, particularly, high concentrations of direct thrombin inhibitors may prolong the assay and thus lower the measured fibrinogen result. In addition, fetal fibrinogen is slightly less reactive to thrombin than adult fibrinogen, and thus measured results may be lower in a clottable assay when used on neonate blood [2]. Dysfibrinogenemia (mutations in fibrinogen) may also lower measured results, as the Clauss method reads only functional fibrinogen.

Factor assays

There are two main forms of factor assays that are performed in coagulation laboratories: clottable, or "one-stage" assays, and chromogenic, or "two-stage" assays. The clottable assays are currently more common, and are called "one-stage" assays as all components are mixed and then immediately tested by either PT or aPTT method, depending on the factor. Chromogenic assays require an activation step, followed by addition of chromogenic substrate for factor Xa. Factor Xa activation is always used as the assay endpoint in chromogenic assays.

In clottable assays, formation of fibrin monomers is the endpoint, the same as in the PT or aPTT assay. However, instead of testing straight patient plasma, the patient's plasma is diluted 1:10 in an isotonic buffer and then mixed 1:1 with plasma that is fully deficient in a single factor. This allows the patient sample to supply the missing factor and complete the coagulation cascade. A PT or aPTT assay is then performed on the mixed sample, depending on the factor being evaluated.

The resulting clotting times can be converted to factor activity using a standard curve, generated with control plasma of known factor concentrations. By convention, a 1:10 dilution (matching the patient sample dilution) of this plasma is set as 100% factor activity, and the plasma is serially diluted to generate the standard curve. The control plasma can also be used in lower dilutions in order to generate a standard curve above 100%. Factor activity is given in percent, and the general conception is that 50% of a factor is sufficient for in vivo clotting; thus normal factor ranges are typically 50%−150%.

In order to exclude the presence of inhibitors, factor assays are typically performed with serial dilutions of patient plasma, such as 1:10, 1:20, and 1:40. If the results of these assays substantially agree after accounting for the

dilution, the results are considered acceptable. However, if the results do not agree, especially if higher dilutions result in higher adjusted factor levels, inhibitors may be suspected.

In addition to one-stage assays, chromogenic assays (two-stage assays) are currently available or in development for many factors, such as FIX, which measure activity through the cleavage of a chromogenic substrate of FXa. In these assays, the patient's plasma is mixed with plasma containing all factors necessary for the activation of FXa except for the factor being tested. The sample is allowed to incubate, and then in the second stage, a solution with chromogenic FXa substrate is added. Activated FXa from the previous step can cleave the substrate, resulting in a chromogenic readout, such as *p*-nitroaniline release, measured by absorbance at 405 nm. The results of these assays should agree with the one-stage assays; however, some cases such as patients on long-acting factor replacements are better monitored through the use of chromogenic assays.

Anti-Xa

A similar assay to the chromogenic factor assays is the anti-Xa assay, commonly used in place of aPTT for heparin monitoring due to the higher specificity and selectivity of the anti-Xa assay. As the name implies, this measures the activity of anti-Xa compounds in the patient's plasma. In this assay, patient plasma is mixed with a solution containing FXa and chromogenic substrate. It may also contain exogenous AT in order to supplement for a patient's potential lack of AT, but formulations differ. Any FXa-inhibiting substances, most typically heparin or LMWH, will inhibit the action of FXa and lower the resulting chromogenic response.

Thrombin time

The term "thrombin time" may be misleading, as the thrombin time is actually independent of endogenous thrombin. It measures the amount of time needed for thrombin to form fibrin strands, and is thus sensitive to the patient's quality and quantity of fibrinogen, as well as to the potential presence of thrombin inhibitors in the sample. The thrombin time is substantially similar to the Clauss method of fibrinogen determination, though the patient sample is not diluted and thrombin concentrations are much lower. Exogenous thrombin, typically bovine or human in origin, and calcium are added directly to a patient's sample, and the time to fibrin strand formation is measured as in the PT or aPTT assays. Low fibrinogen and/or dysfibrinogenemias (mutations in fibrinogen) will prolong the thrombin time. Any inhibitor of thrombin will also prolong the thrombin time, such as heparin or direct thrombin inhibitors (e.g., dabigatran). Notably, LMWH will typically not prolong a thrombin time, as it does not inhibit thrombin; Xa is bypassed in this assay. In addition, elevated fibrin, FDPs, and paraproteins can prolong the thrombin time through interference of thrombin activity or fibrin strand formation. Fetal fibrinogen may also delay clotting times in a similar manner to the fibrinogen assay.

D-dimer

After fibrin monomers are formed, degradation of the solidifying clot by plasmin will result in FDPs, also known as fibrin split products. While assays exist for general measurement of FDPs, this has mostly given a way to the testing of D-dimers, which are a specific subset of FDPs that include cross-linked fibrinogen D domains. Fibrinogen D domains are specifically cross-linked by FXIIIa, and thus D-dimers are only produced in a mature clot. FDPs, however, can be formed by any fibrolytic event, and elevated concentrations are observed in many conditions including liver disease, eclampsia, cirrhosis, and acute thrombotic events. D-dimer is considerably more specific to thrombotic events, and is the preferred analyte to measure if a patient is under suspicion of a thrombotic event.

The D-dimer assay is also useful in the background of pulmonary embolism (PE) and deep vein thrombosis (DVT), where D-dimer has an excellent negative predictive value. The positive predictive value is marred by many possible causes of fibrinolysis such as trauma and recent surgery; however, the D-dimer is a very effective rule-out test for PE and DVT in emergency settings, with negative predictive values of 95%−100% [3,4]. For PE cases with a low pretest probability (which can be calculated with many tools such as the Wells score [5], a method for quantifying PE risk based on clinical features), a negative D-dimer assay can save a patient and care team the time, expense, and radiation exposure of a CT scan.

D-dimer products are typically measured by latex immunoagglutination (LIA) assays. LIA assays are a form of immunoassay, where the antibodies are linked to latex beads. Recognition of the antigen by the antibodies causes agglutination of the beads, and this causes increased turbidity and decrease of absorbance of the sample which can be read by an analyzer. D-dimer units (DDUs) of reporting can cause a great deal of confusion, as they are not standardized across the industry. In general, two units are used: the fibrinogen equivalent units (FEUs) or DDUs. DDU is the simpler unit, as it is only the microgram per milliliter of D-dimer detected in the plasma. FEU, alternately, relates the amount of D-dimer to the amount of fibrinogen that the D-dimer was originally; thus it is 2xDDU. A given manufacturer or reference lab

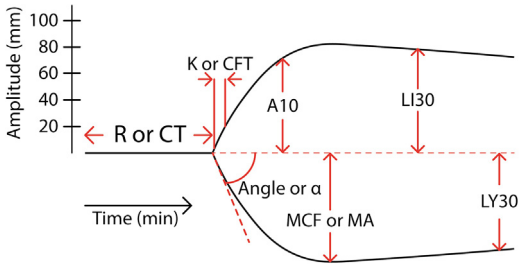


FIGURE 26.8 Typical thromboelastography graph. Parameters for both thromboelastography and rotational thromboelastometry are shown. R or CT is the time for the amplitude to reach 2 mm, and is analogous to the prothrombin time or activated partial thromboplastin time, depending on clot activator used. K or CFT is the time required for the amplitude to increase from 2 to 20 mm. Angle or α is the angle tangent to the curve at the start of clot formation. A10 (rotational thromboelastometry) is the amplitude 10 minutes after CT. The MCF or MA is the maximum "strength" of the clot, and is the maximum amplitude of the graph. LI30 (rotational thromboelastometry) is the reduction in strength of the clot 30 minutes after CT. LY30 (thromboelastography) and the associated LY60 (not shown) are the percent reductions in clot strength 30 and 60 minutes after MA.

may use DDU or FEU, and this should be checked carefully before using the results and should be clear on any results put out by a laboratory.

Thromboelastography

Thromboelastography (TEG) is a whole-blood viscoelastic test, which attempts to capture a large degree of information about a patient's bleeding status. Unlike standard laboratory tests for coagulation, TEG is performed in the presence of platelets as well as other blood factors that may affect clotting typically removed by centrifugation. In addition, as there is no centrifugation required, this technique can be used in a close-to-bedside setting. However, TEG is a nonwaived test and is performed frequently by laboratory personnel. There are two main devices that are used for this purpose in laboratory settings: the TEG (Haemonetics, Braintree, MA, USA) and rotational thromboelastometry (ROTEM; Instrumentation Laboratory, Bedford, MA, USA). These devices work under the same principle: a sample of anticoagulated whole blood is placed in a heated circular cup with a pin in the center. The sample is warmed, calcium and clot activator are added, and then either the pin (ROTEM) or cup (TEG) is rotated back and forth to generate a low shear stress. The force applied on the pin by the forming clot is translated into an amplitude, which is then plotted, and several different factors calculated from this graph (Fig. 26.8).

The CT (ROTEM) or R (TEM) is analogous to the PT or aPTT, depending on the clot activator used. Beyond this, various factors calculate the overall clot strength, which is dependent on fibrin strength and platelet activation, and fibrinolysis. The results of TEG can be used to help guide blood component replacement, and several studies have shown its utility in decreasing blood component cost through more specific direction in settings such as cardiac surgery [6,7].

There are many calculated variables and modifications to TEG, which purport to give additional information on hemostasis and alter the testing time necessary. The breadth of these variables and modifications is outside the scope of this work. For more information, please see various reviews such as "TEG and ROTEM: technology and clinical applications" by Whiting and DiNardo in the *American Journal of Hematology*.

Pathophysiology and advanced coagulation testing

Hemostasic disorders can be broadly categorized as either increased bleeding or increased clotting (thrombophilia). Bleeding disorders are predominantly caused by von Willebrand disease (vWD) with specific deficiencies observed less frequently, while thrombophilias are a more diverse group that may be caused by autoimmune inhibitors, exogenous inhibitors, and deficiencies. Most testing for these disorders is typically termed "advanced coagulation testing" or "special coagulation testing," and may not be performed in a core laboratory.

Bleeding disorders

Von Willebrand disease

As previously mentioned, bleeding disorders are dominated by vWD, which is itself a cluster of different disorders based around the functioning of vWF. Through testing of the vWF activity, antigen amount, FVIII activity, multimer pattern, and platelet response to ristocetin, these disorders can be distinguished from one another (Table 26.1). A typical workflow for suspected vWD includes testing vWF antigen (a mass assay, typically by immunoassay) and activity along with FVIII activity. Those with suspected Type 2 vWD should have multimers tested, followed by ristocetin-induced platelet aggregation (RIPA) testing in those with suspected Type 2B [8,9]. RIPA testing is described in more detail in the platelet aggregation section, below.

Ristocetin is commonly used in assays for vWF activity. Originally utilized clinically as an antibiotic, it was removed from the market due to causing thrombocytopenia. Ristocetin facilitates the GP Ib/vWF interaction, which causes platelets to be activated and then removed from circulation. In the ristocetin cofactor assay, patient's plasma is mixed with exogenous GP Ib (typically in the form of normal platelets or GP Ib-coated beads) and

TABLE 26.1 Typical results for vWD tests for all major subtypes.

vWD type	vWF antigen	vWF activity	FVIII activity	Multimer pattern	RIPA response
Type 1	Low	Low	Low	All present at lower concentrations	Normal response
Type 1C[a]	Low	Low	Low	Often has ultra high-molecular-weight bands present	Normal response
Type 2A	Normal	Low	Normal	Loss of high-molecular-weight bands	Reduced response at 1 mg/mL ristocetin
Type 2B	Normal	Low	Normal	Loss of high-molecular-weight bands	Increased response at 0.5 mg/mL ristocetin
Type 2M	Normal	Low	Normal	Normal pattern or increase in high-molecular-weight bands	Reduced response at 1 mg/mL ristocetin
Type 2N	Normal	Normal	Low	Normal pattern	Normal response
Type 3	Very low	Very low	Very low	No bands	Reduced response at 1 mg/mL ristocetin

RIPA, Reistocetin-induced platelet aggregation; *vWD*, von Willebrand disease; *vWF*, von Willebrand factor.
[a]*Type 1C will show an elevated vWF propeptide-to-antigen ratio where this testing is available.*

ristocetin. If the vWF in the patient's plasma is normal, ristocetin will facilitate the vWF binding to GP Ib and cause agglutination, which can be measured optically. If the patient's vWF is unable to bind GP Ib, agglutination will be highly reduced or absent, causing a lower reading.

VWD can be the result of either a quantitative loss of vWF or a qualitative dysfunction. It results in inadequate primary hemostasis, which may manifest as epistaxis, heavy menstrual bleeding, and easy bruising. The condition can be associated with the decreased antigen and functional activity (Types 1 and 3) or with a relatively normal antigen concentration and reduced functional activity (Type 2). vWD, as defined by laboratory testing, has an overall prevalence of 8000 per million population, while clinically significant disease with bleeding affects about 100 individuals per million population. The most common quantitative disorder, Type 1, can occur for many reasons and constitutes 90% of all vWD. Known causative mutations are found in ~65% of cases, and are an autosomal dominant condition. It is diagnosed as vWF activity and antigen below 50% of normal serum, and is also associated with low FVIII activity. This can come about through missense, nonsense, and promoter mutations in the vWF gene, as well as through consumption. As vWF is activated and exposed to ADAMSTS13 in areas of high shear stress, introducing more stress can reduce levels of active vWF such as in ECMO. This is also termed "acquired Type 1 vWD." Other sources of acquired Type 1 vWD include monoclonal gammopathies, non-Hodgkin's lymphoma, myeloproliferative disorders, some solid tumors, hypothyroidism, sodium valproate, and ciprofloxacin treatment, among others. Type 1 vWD is both the most common (up to 2% of the population is affected) and one of the most diverse bleeding disorders from a causative standpoint.

In comparison with the classic Type 1 vWD, there is an alteration in the clearance known as "Type 1C vWD." Type 1C is characterized by a significant rise in vWF clearance, resulting in low vWF levels and increased responsiveness to desmopressin (DDAVP); however, there is also a significantly shortened half-life for DDAVP treatment. Patients with Type 1C are identified by an abnormal vWF-propeptide to vWF antigen ratio (elevated in patients with Type 1C vWD). Type 1C can also be found through a time course study, which is typically performed on new DDAVP patients. Type 1C patients will show normal or high response to DDAVP (seen as increase in vWF antigen and activity 1 hour after administration), followed by a significantly higher reduction at 4 hours after administration. This has therapeutic implications, as DDAVP is typically used to prepare Type 1 patients for minor surgery and may not be suitable for Type 1C patients. It should be stated that individuals with blood group "O" are known to have lower vWF levels in general, most likely through an increased clearance. The significance of this decrease in levels to the diagnosis of vWD is debated, and symptomatic correlation is suggested in these patients [10].

Type 2 vWD is defined as a qualitative loss of vWF activity. This can come about through various causes, which breaks down the different subtypes. In almost all cases, the defining feature is a discrepancy between the activity and antigen results. In most cases, the vWF activity to antigen ratio should be greater than 0.6. In Type 2

vWD, the antigen will be normal or slightly decreased, while the activity will be much lower. This is true for types A, B, and M. Another informative test in vWD Type 2 is to examine the state of multimerization of vWF. Multimer patterns are typically assessed by electrophoresis on agarose gels followed by Western blot analysis with a specific antibody to vWF. Normal plasma produces an extended ladder of multimers, including high-molecular-weight forms.

In Type 2A vWD, there is a mutation in the ADAMTS13 cleavage site which causes increased cleavage. These mutations are of various types and at various locations on the vWF sequence, leading to either a dominant or recessive phenotype. In all cases, vWD Type 2A results in a loss of high-molecular-weight vWF. This also leads to a lack of response of platelets to ristocetin, which can be seen in platelet aggregation studies.

Type 2B is from a gain of function mutation in the GP Ib/IX/V binding site. This causes vWF to be inappropriately bound to platelets and removed from circulation. This results in the loss of very high MW vWF multimers as well as mild thrombocytopenia, as bound platelets are also removed from circulation. Platelets from these patients are also more sensitive to low-dose ristocetin.

The opposite of Type 2B, Type 2M, is the result of a loss of function mutation in the GP Ib/IX/V binding site on vWF. As the multimers do not bind platelets well, this leads to either a normal distribution of multimers or even an increase in high-molecular-weight multimers. The response to ristocetin by these platelets is muted, similar to Type 2A.

The last Type 2, Type 2N, is the result of mutations involved in the binding of FVIII to vWF. As a result, FVIII cannot bind to vWF, and free FVIII is cleared quickly from the circulation resulting in deficient FVIII activity levels though vWF activity, antigen, and multimers are all typically in the normal range. There is also no effect on a patient's platelet's response to ristocetin.

The last type of vWD, vWD Type 3, is a more severe form of Type 1 and results from mutations that lead to a complete loss of vWF. vWF activity, antigen, and FVIII activity are all significantly decreased to immeasurable, and bleeding is typically severe and can easily occur in both mucosa and soft tissues. Without treatment, vWD Type 3 is fatal.

Hemophilia A and B

Hemophilia A and B are the results of single-factor deficiencies, either FVIII (hemophilia A) or FIX (hemophilia B). They are often inherited as X-linked disorders and affect up to 1 in 5000 patients; as many as 30% of all cases are the result of de novo mutations. Hemophilia A is the most common and accounts for approximately 85% of cases, and hemophilia B accounts for approximately 14% of cases. The remaining 1% of cases are the result of other deficiencies, such as FXI, and are exceedingly rare. The exception to this is FXI deficiency in Ashkenazi Jewish populations, where the incidence can rise as high as 4%. However, FXI levels are only weakly correlated with bleeding.

Hemophilia A and B have the same classification and symptoms. Both conditions can be classified into three distinct categories based on their factor activities: (1) factor activity between 6% and 20% is "mild"; (2) factor activity between 1% and 5% is "moderate"; and (3) $<1\%$ is "severe." Bleeding is directly correlated with factor levels, and factor replacement is often used to treat hemorrhage or to prepare for invasive surgery. Severe forms may require chronic infusions. Treatment of these conditions can also be complicated by the development of autoimmune factor inhibitors. Because of the constant use of exogenous factors, up to 30% of severe hemophilia A patients and 2%–3% of hemophilia B patients develop these alloantibodies after several doses, which can be determined through a lower than expected response to factor infusion as well as through an inhibitor assay.

In rare cases, a patient may develop a spontaneous autoimmune FVIII inhibitor. This is termed "acquired hemophilia" and affects one in a million patients. Clinical symptoms are severe and mimic severe hemophilia A, though it typically occurs in elderly patients and frequently with comorbid autoimmune conditions such as lupus or rheumatoid arthritis. Similar to the alloantibodies mentioned previously, an autoimmune inhibitor can be detected through an inhibitor assay. A mixing study will also reveal the presence of an inhibitor, though a specific inhibitor study is needed to determine severity.

Factor deficiencies

Congenital factor deficiencies beyond VIII and IX are extremely rare. Specific acquired deficiencies are also quite rare but do occur more often, and the inhibitors that are at fault in these cases can be determined through an inhibitor assay in the same manner as acquired hemophilia, using the affected factor assay as a basis. The original deficiency can be found with factor assays, with inhibitor assays utilized after the specific deficiency is pinpointed.

As mentioned previously, both liver and vitamin K status can profoundly influence the coagulation cascade. Both will extend the PT, and severe forms of both will prolong the aPTT. However, liver disease will have a more pronounced effect on the aPTT than that of the PT. A typical response to an isolated prolonged PT, especially without overt bleeding symptoms, will be a two-factor

test—usually FV and FVII or FV and FX. As FV is vitamin K-independent, both tests will only be low in the case of liver disease. If only FVII or FX is low, vitamin K deficiency is more likely.

Platelet disorders

Platelets are composed of many highly specialized receptors and organelles, and congenital or acquired disruption of these can cause many bleeding disorders, though most present similarly as mucocutaneous bleeding of various severities and may or may not present with thrombocytopenia. Many of these disorders or deficiencies can be determined through the use of platelet aggregometry, which measures both the ability of platelets to aggregate as well as the ability of the platelets to secrete. Platelet aggregometry can be performed on either platelet-rich plasma or whole blood. Aggregometry performed on plasma uses light transmission for detection, which increases upon platelet aggregation, while whole blood uses electrical impedance, which also increases upon platelet aggregation. Both whole blood and plasma aggregometry typically include the use of the luciferase enzyme that causes a measurable output of light when platelets release ATP from their dense granules. A normal control is always used for comparison to the patient sample, and then both pools of blood or plasma are challenged with various agonists to induce aggregation or granule release. Typical agonists include thrombin, AA, ADP, collagen, and ristocetin. Epinephrine may be used, though this is only utilized in PPP.

Most agonists produce platelet aggregation through a GP IIb/IIIa-dependent mechanism with the notable exception of ristocetin. RIPA depends on the presence of both vWF and the GP Ib/IX/V complex, and is independent of the GP IIb/IIIa receptor. RIPA is reduced or absent in most forms of vWD, whereas ADP, epinephrine, AA, and collagen produce normal aggregation responses in this condition. The exception is vWD Type 2B, which shows an increased response to low-dose ristocetin. Various platelet disorders can be determined through the particular pattern of aggregation and release that each disorder shows. Also, microscopy can help to confirm the diagnosis, especially in the cases of storage pool disorders.

Acquired platelet disorders

Many conditions may cause platelet function defects, either congenital or acquired. Acquired disorders are more common and have a staggeringly large number of sources. For this section, we will focus only on the most common.

By far, the most common cause of acquired platelet dysfunction is aspirin use. Aspirin irreversibly inhibits COX proteins, leading to a profoundly impaired or even absent aggregation and release in response to AA, the substrate of COX-1, as well as reduced response to ADP and collagen. In whole blood, ADP may still show a normal aggregation response. However, thrombin, which activates platelets independently of COX-1, will show a normal response. A failure to aggregate or secrete with all of these agonists with preserved thrombin response may be attributable to aspirin use and may suggest a retest if a platelet function disorder is still suspected.

Most nonsteroidal antiinflammatory drugs (NSAIDs), which likewise function as COX inhibitors, also impair platelet function. However, as these are not irreversible inhibitors, the time frame of inhibition is much shorter. Typical doses of other NSAIDs, such as naproxen sodium or ibuprofen, may show selective inhibition of platelet activity for up to 3 days. While not typically classed as an NSAID, acetaminophen can also affect platelet function with a similar pattern. It has an even more reduced effect, however, beginning to correct within 90 minutes of administration. As such, inhibition by acetaminophen should only be considered if use is suspected very close to the time of draw.

Clopidogrel is another antiplatelet drug that is commonly used alongside aspirin and inhibits the $P2Y_{12}$ receptor, which is targeted by ADP. As such, response to ADP is reduced. Unlike aspirin, the $P2Y_{12}$ effect is limited to ADP and collagen, which will both show a reduced response in terms of aggregation and secretion.

Congenital platelet disorders

Most platelet aggregation is done to diagnose congenital disorders, which can take many forms. Lack of α granules, dense granules, deficiency of secretion enzymes, and membrane receptor defects are all described and have different patterns of aggregation, release, and microscopy (Table 26.2).

Deficiency of α granules is also known as "gray platelet syndrome," as alpha granules give platelets their distinctive speckled appearance on light microscopy and the platelets are too large. Since α granules contain various coagulation factors as well as vWF, lack of α granules results in a muted response to all agonists.

Deficiency of dense granules has a similar effect, though this will not be visible using light microscopy. Electron microscopy is capable of imaging dense granules and can illustrate their absence. Platelet aggregation will show a similarly muted response to all agonists, especially in terms of granule release, which is measured based on dense granule ATP release. If this defect is suspected, electron microscopy can be performed, which can determine if dense granules are decreased relative to normal patient controls. When dense granules are

TABLE 26.2 Aggregation results for various platelet disorders when tested by platelet aggregometry.

Disorder	ADP	Epinephrine	AA	Collagen	Thrombin	Ristocetin
Aspirin or aspirin-like	Reduced	Reduced	Reduced	Reduced	Normal	Normal
Bernard–Soulier	Normal	Normal	Normal	Normal	Normal or reduced	No response
Glanzmann's thrombasthenia[a]	No response	Reduced	No response	No response	Reduced	Normal
Storage pool disorder[b]	Reduced	Reduced	Normal or reduced	Reduced	Reduced	Normal
Membrane defect	Reduced	Reduced	Normal	Reduced	Normal	Normal

AA, Arachidonic acid.
[a]*Glanzmann's thrombasthenia will typically show normal ATP release by platelets with all agonists.*
[b]*Dense and α granule storage pool disorders can be distinguished by microscopy. Loss of α granules results in large, "gray" platelets (no visible granules) on light microscopy. Loss of dense granules can be seen in electron microscopy.*

decreased in patients with oculocutaneous albinism, Hermasky–Pudlak or Chediak–Higashi syndrome should be suspected. In those males with decreased dense granules and without oculocutaneous albinism, Wiskott–Aldrich syndrome is most likely.

Defects in COX enzymes can also occur, which will result in lab results that are highly similar to those of aspirin treatment, and may be called an "aspirin-like defect." In this case, very careful patient histories and repeat testing may be required to distinguish the disorder from a drug effect.

Platelet type or pseudo-von Willebrand disease (pseudo-vWD) is a platelet receptor defect disorder that causes a similar clinical and laboratory picture as vWD Type 2B. This does not involve any mutation on vWF but is instead a mutation of the GP Ib/IX/V receptor. A gain of function mutation in GP Ib increases affinity for vWF in a similar manner to vWD Type 2B, resulting in a similar clinical picture. Testing will also show similar results, with a loss of high and intermediate size vWF multimers and increased platelet response to ristocetin. One method to distinguish the two is through the addition of vWF to patient platelets, typically through the use of cryoprecipitate. If a patient has vWD Type 2B, the addition of cryoprecipitate will at least partially correct the response to ristocetin. In pseudo-vWD, the error is on the platelets, and thus the addition of vWF will not correct the results [11].

Much as pseudo-vWD may masquerade as vWD Type 2B, Bernard–Soulier syndrome (BSS) may masquerade as vWD Type 2M. This disease is the result of a decrease or loss of the GP Ib/IX/V complex on the surface of platelets. Multiple different mutations may cause this disorder, spread across the three component proteins. All of these mutations result in the same features of thrombocytopenia as well as characteristic large platelets, as the GP Ib/IX/V complex is also required for proper production of platelets by megakaryocytes. In platelet aggregation studies, BSS patients will react similarly to vWD Type 2M and show a normal response to almost all agonists except for ristocetin, which will display a markedly reduced or absent response. BSS and vWD Type 2M can also be distinguished through the use of exogenous vWF. As the platelets are affected in BSS, addition of vWF will not correct the response to ristocetin, while in vWD Type 2M, the addition will correct the response.

Glanzmann thrombasthenia (GT) is the result of a defect in the GP IIb/IIIa platelet surface receptor. GT is characterized by prolonged or spontaneous bleeding from birth, presenting as easy bruising, gum bleeding, and epistaxis, as well as petechiae and hematomas. As in BSS, the causative mutations vary and can occur on either the IIb or IIIa protein. The GP IIb/IIIa receptor is required for platelets to bind fibrinogen; thus the formation of clots is highly attenuated in this condition. As a result, platelet aggregometry will show little to no aggregation for all agonists except ristocetin, which may show aggregation as platelets bind to vWF. However, fibrinogen binding is required to form a mature clot, and so disaggregation is sometimes seen with ristocetin aggregation. Also, all reagents will typically show degranulation and ATP release, as GP IIb/IIIa is not required for degranulation.

Testing for bleeding disorders

Most patients with bleeding are tested with PT and aPTT and a complete blood count when presenting to a hospital, and these results can help guide further testing. Isolated increases should be followed by mixing studies. Correction in the mix indicates a factor deficiency. Isolated PT prolongation or prolongation in both with correction should be followed by clinical correlation and factor assays to assess liver function and vitamin K status. Isolated aPTT prolongation with correction should be followed by testing for FVIII inhibitors and hemophilia. Correction on mix but prolongation on incubation is

highly indicative of an FVIII inhibitor (acquired hemophilia A).

Bleeding in the absence of PT or aPTT prolongation, especially with normal platelet counts, should be followed by vWD testing, which should be both activity and antigen testing to determine vWD Type 1 (concordant low results), 2 (discordant results), or 3 (concordant severely low results). Suspicion of Type 2 should prompt multimer and RIPA testing, if available, to determine subtype. Mucocutaneous bleeds, especially in the setting of normal PT, aPTT, and vWD testing, should prompt platelet aggregation testing to isolate a platelet function disorder if drug effect can be ruled out. Abnormal platelet aggregation should also be followed by light and electron microscopy if the results are consistent with a storage pool disorder.

Thrombophilias

Thrombophilias are a diverse group and are defined primarily by an increased risk for the formation of a venous thromboembolism (VTE). VTEs include both deep-vein thrombosis as well as PE. This increased risk can come about through either hyperactivation of the coagulation pathway or a lack of proper controls. VTEs are also distinct from arterial thrombosis, which are typically caused by cardiovascular or cerebrovascular issues. The acquired causes are by far the most common, and while there is a tendency to focus on those thrombotic conditions for which well-established tests exist, it is worthwhile noting that the majority of acquired thrombotic disorders are not amenable to laboratory testing. These include:

- inflammation, trauma, and postsurgical states;
- venous stasis and limb immobilization;
- estrogen-containing oral contraceptives and pregnancy; and
- malignancy.

One acquired cause for which laboratory testing exists is the antiphospholipid syndrome (APS; see below). Inherited causes of thrombophilia are relatively uncommon, although it is essential to recognize their existence by the presence of thrombosis at a young age, a family history of thrombosis, and thrombosis at unusual sites.

Activated protein C resistance

The most common congenital thrombophilia, especially among Caucasians, is activated protein C resistance (APCR), though it may also be acquired. In short, any mechanism that decreases the ability of APC (i.e., PC bound to its cofactor PS) to inactivate FVa and FVIIIa can lead to APCR. This can be diagnosed through the simple use of the addition of APC to a serum sample. In healthy individuals, this should prolong the PTT, as the exogenous APC will block the formation of prothrombinase. In patients with APCR the PTT will not prolong as the exogenous APC will be ineffective. A ratio of the PTT results with and without APC can be used to determine the extent of the effect. This is also known as the "APCR-standard assay" (APCR-S).

A modified APCR (APCR-M) assay is also available, and is based on the PTT clotting time using plasma treated and untreated with APC and with the addition of factor V depleted plasma—this is also expressed as a ratio of the clotting times similar to that of APCR-S. However, the APCR-S assay is affected by anticoagulation therapy, and is less specific and sensitive for factor V Leiden (FVL) than the APCR-M assay. Although the APCR-S assay has higher sensitivity for acquired forms of APCR, such as seen in pregnancy, use of oral contraceptives, and in patients with APS than the APCR-M assay, the APCR-M assay is less influenced by factor deficiencies, oral anticoagulation, and lupus inhibitors.

The most common congenital form of APCR is the FVL mutation, which occurs in approximately 8% of patients of European decent and up to 90% of patients with APCR. The FVL mutation is specifically the G1691A (DNA), or R506Q (amino acid), mutation of the FV gene. This renders it highly resistant to proteolysis by APC leading to prolongation of coagulation as well as spurious activation. Homozygous patients for FVL are up to 18 times more likely than the general population to develop a thrombotic event, though heterozygous patients are only three times more likely. The APCR-M assay is 99.6% sensitive and 99.7% specific for the FVL mutation. In addition to FVL, there are several other FV mutations and FVIII mutations that can lead to APCR, though reports on the penetrance of these mutations are uncertain. Also, there are reports of an increase serum concentration of FV, FVIII, and FIX in acquired APCR patients, though the significance is currently unknown [12].

Prothrombin G20210A

Prothrombin G20210A (PGM) is a polymorphism in the prothrombin (factor II) gene that is associated with an increased risk of VTE. The carrier frequency is approximately 1.5%–3% in Caucasians. Heterozygous individuals for the mutation have approximately a fivefold increased risk of VTE. The mutation, in the 3' untranslated region of the mRNA, does not affect the protein structure of prothrombin, and the amino acid sequence is unchanged. The mechanism of action of this polymorphism is not entirely clear, but it appears to be related to increased plasma concentrations of prothrombin (factor II). Currently, there are no functional assays for PGM,

and the molecular testing methods available are similar to those available for FVL.

Protein C and S deficiency

The anticoagulant PC and its necessary cofactor PS may each be decreased in patients, which leads to a similar clinical picture of FVL in severe cases. However, these deficiencies will appear only as a low PC or PS upon analysis, and will not show positive results in an APCR assay. Congenital deficiencies are quite rare (<1% of the general population), and are almost always heterozygous, as the homozygous form can be fatal in utero. Those that survive may suffer diffuse intravascular thrombosis as well as massive venous thrombosis. In these patients, either PC or PS will be undetectable. This must be treated with oral anticoagulants, PC or PS replacement, or liver transplant, and the last of which is very rare [13].

Acquired deficiencies are much more common, and have a range of causes. As PC and PS are both manufactured by the liver and vitamin K-dependent, liver failure/insufficiency and vitamin K antagonists are both common causes of acquired deficiency. Sepsis has also been shown to cause PS and PC deficiencies, and PS decreases during pregnancy.

Protein C

Testing for PC in plasma is performed using either a clottable or chromogenic assay. In the clot-based PC functional assay, the patient plasma specimen is mixed with a commercial PC-deficient plasma. A PC activator, usually a snake venom derived from *Agkistrodon contortrix* (copperhead snake) is added to the mix. This will convert the endogenous PC to activated APC. When clotting is initiated through a PT- or aPTT-based reaction, the clotting time is proportional to the endogenous functional PC concentration. The higher the PC, the more the FV and FVIII are inactivated and the more the clotting time is prolonged. The clot-based assays provide some advantages: (1) sensitive to Type I (quantitative loss) and Type II (qualitative loss) PC deficiency; and (2) increased sensitivity for most coagulation instruments, regardless of fibrin clot detection methods. Some of the disadvantages include (1) underestimating PC when FVIII is elevated (>250%), PS is low in the test plasma, FVL is present, warfarin antagonism, or vitamin K deficiency; and (2) overestimation of PC if the test plasma contains heparin, DTI's, anti-Xa, lupus inhibitor, or specific factor inhibitors. An alternative functional PC assay is the chromogenic or amidolytic PC assay. In this assay, the endogenous PC is activated through the action of *Agkistrodon* venom, and the APC is detected by its ability to cleave a synthetic substrate with the release of *p*-nitroaniline. This reaction is unaffected by heparin, antiphospholipid antibodies (APAs), anticoagulants, FVIII, and FVL.

Protein S

As indicated above, PS together with APC binds to the phospholipid surface of platelets or other cells, enhancing the enzymatic inactivation of FVa and FVIIIa. Approximately 40% of PS is found free in plasma, and the remaining 60% is bound to complement system protein C4b-binding protein (C4bBP). Different pathological and physiological conditions can alter the ratio of bound to free PS, thus altering the PC/PS inhibitory pathway. For example, studies have shown that during pregnancy, there is a decrease in the level of PS, with values averaging only 60% of those of nonpregnant controls. Also, acquired PS deficiency can be seen, when there is a decrease in synthesis (hepatic disease, vitamin K antagonists), increase in clearance (acute thrombosis, nephrotic syndrome), and hemodilution. Furthermore, genetic deficiencies are classified into Type I or quantitative defect, Type II or qualitative defect, and Type III or increased binding affinity to C4bBP.

In the functional PS assay, patient plasma is mixed with a commercial PS deficient plasma, and an exogenous PC is added. This will cleave FVa and FVIIIa, provided endogenous PS is present. Once again, clotting is initiated through the PT or aPTT, and the clotting time is proportional to the PS concentration. In some of these functional assays, exogenous factor V is added to the reaction mix; if the patient is heterozygous or homozygous for FVL, the reaction will show increased resistance to APC, which will manifest as a falsely decreased PS. Addition of exogenous normal FV will prevent this problem. Furthermore, very high factor VIII concentrations (typically >250%) can also shorten the clotting time, again leading to a falsely decreased S. In some reaction setups, the factor VIII step is bypassed using activated factor Xa to initiate the common pathway. Any effects that prolong clotting times can falsely elevate the PS in these assays, including very high heparin (>1.0 U/mL), direct thrombin inhibitors including the new oral anticoagulants, and APAs. Antigenic assays (for free and total PS antigen) are available and are generally considered to be more reliable.

Antithrombin deficiency

Much like PC or PS deficiency, congenital defects of AT are rare, also at less than 1% prevalence. AT deficiencies are always heterozygous, as homozygous mutations are fatal in utero. Heterozygous AT deficiency can lead to an approximately 15-fold increase in thrombosis risk, similar to that of homozygous FVL. Two types of congenital AT deficiencies are noted. Much as in other disorders, Type I is a quantitative loss of AT, and Type II is a qualitative loss. Type II can be further subcategorized into (1) reactive site mutations (altering AT interaction with factor Xa and thrombin active

sites); (2) heparin binding site mutations (mutations that affect the binding of heparin to AT); and (3) pleiotropic mutations (mutations that affect multiple heparin binding sites and reactive sites).

Acquired AT deficiency is more common. AT is not vitamin K-dependent, but as it is manufactured by the liver, failure or insufficiency will cause a deficiency, as well as disseminated intravascular coagulation (DIC), nephrotic syndrome, and, during the acute period, a venous thrombosis (consumption). In hospital settings, however, a common cause of acquired AT deficiency is the use of high-dose heparin in settings where consumption is occurring, such as in the case of ECMO patients. Activation of AT by thrombotic events, such as passage through extracorporeal tubes and subsequent binding by heparin, results in an increased clearance from the blood and a drop in serum levels. As heparin requires AT to function, this can likewise result in a similar drop in heparin effectiveness. If not adequately monitored, this can result in catastrophic thrombosis, as both AT and heparin cease to function. In cases of heparin use where AT is falling, supplementation with AT infusion is required if crossover to a different anticoagulant is not possible.

AT functional assays are all chromogenic with cleavage of a synthetic substrate by either thrombin or FXa—both formats are available. The reaction takes place in the presence of heparin that is supplied in the reagent. This heparin will bind to the AT from the patient specimen and will inhibit the amidolytic release of the chromogen. Note that these assays can be affected by the new oral anticoagulants. Dabigatran will inhibit the thrombin step, while the direct Xa inhibitors (i.e., rivaroxaban) will inhibit Xa. Both lead to a falsely elevated AT activity. AT antigen assays are performed through immunoassay.

Plasminogen deficiency

Plasminogen is the major regulator of fibrinolysis, and though rare, both qualitative (Type 1) and quantitative (Type 2) deficiencies exist. A deficiency can be suspected in cases where a patient presents with a thrombotic event despite no evidence of a hypercoagulable state such as natural anticoagulant deficiency, or in cases of monitoring for thrombolytic therapy such as anistreplase. Plasminogen can be tested for by either an antigen assay or an activity assay. The antigen assay is an immunoassay, while the activity assay is chromogenic, involving a specific plasmin substrate. A typical algorithm is to test first with activity, and then only test antigen if the activity is low, as normal activity means that a deficiency is not present. Plasminogen is elevated in pregnancy and with the use of oral contraceptives, and is a positive acute phase protein. Plasminogen is synthesized in the liver and thus is sensitive to liver disease, and is also low in newborns up to 6 months of age. Plasminogen can also be lowered temporarily by consumption; therefore DIC, thrombolytic therapy, and other thrombolytic states will show a loss.

Heparin-induced thrombocytopenia

Another relatively rare but hospital-relevant condition is heparin-induced thrombocytopenia (HIT). In around 10% of cases of heparin administration, antibodies are generated against the heparin-PF4 complex. In up to half of these cases (5% of cases of heparin administration), the resulting complex activates platelets, causing severe thrombocytopenia as well as life-threatening thrombosis. This can also occur in cases of LMWH administration, though the incidence is much lower. HIT typically occurs at least 5 days after the start of heparin administration, though it can occur within a day if there was previous exposure to heparin within 100 days prior.

As a result of this possibility, it is recommended that patients on heparin therapy have their platelets monitored serially, starting before therapy initiation. The degree of thrombocytopenia, the onset of thrombocytopenia, occurrences of new thrombi, and other causes of thrombocytopenia can all be used to grade the possibility of HIT through the use of the 4T score, a common tool for clinical assessment (Table 26.3) [14]. In all cases of suspected HIT, heparin therapy should be discontinued and replaced with a nonheparin anticoagulant, as LMWH can also cross-react with anti-PF4-heparin antibodies [15].

Screening for suspected HIT can be performed using an immunoassay against the PH4-heparin complex. The confirmatory and reference method is the serotonin release assay (SRA), which measures serotonin release from patient platelets in the presence of heparin. Patient platelets are mixed with platelets labeled with radioisotope-labeled serotonin, and the release is then monitored in the presence of various levels of heparin In an HIT patient, the labeled platelets will release serotonin in the presence of moderate amounts of heparin (0.1−0.3 IU/mL), but not with high levels (100 IU/mL). Activation of both moderate and high levels of heparin reveals a nonspecific reaction, as high heparin levels will disrupt the causative PF4-heparin complexes. When performed correctly and in the presence of a positive screen and moderate or high T4 score, the SRA is highly specific for HIT [16].

Antiphospholipid syndrome

The most common acquired thrombophilia is APS. As the name implies, these are autoantibodies against phospholipid-binding proteins, such as β_2-glycoprotein 1, known as APAs. These antibodies are frequently referred to as LAs, though they are prothrombotic and less than half of the cases of LAs occur in patients with systemic lupus erythematosus (SLE). However, patients with SLE and other autoimmune diseases are much more likely than

TABLE 26.3 The 4T score.

T category	2 points	1 point	0 points
Thrombocytopenia	Platelet drop ≥50% to a low of ≥20,000/mL	Platelet drop 30%–50% to a low of 10,000 to 19,000/mL	Platelet drop <30% or low of <10,000/mL
Timing of platelet drop	Clear onset 5–10 days after first administration, or within 1 day if previous heparin therapy within the last 30 days	Unclear onset within 5–10 days, clear onset ≥10 days after, or fall within 1 day with previous heparin therapy 30–100 days prior	Platelet count drop <5 days without recent exposure
Thrombosis or other symptoms	New, confirmed thrombosis, skin necrosis, and/or acute systemic reaction postheparin administration	Progressive, recurrent, or suspected thrombosis or nonnecrotizing skin lesions	None
Other causes of thrombocytopenia	None apparent	Possible	Definite

Notes: The 4T score is an effective method of stratifying risk for heparin-induced thrombocytopenia. A low total 4 T score (1–3) has high negative predictive value, and thus the 4 T score is a good rule-out test. Intermediate (4–5) and high (6–8) scores should be followed up with PF4 and SRA testing.

the general population to develop APAs, with 30% to 50% of SLE patients developing APS. When a diagnosis of APS is made in the absence of another autoimmune condition, it is known as primary APS. Otherwise, it is considered secondary APS. Only 8% of cases of primary APS will proceed to develop an additional autoimmune diagnosis.

Laboratory testing of APS is rather diverse, which reflects the diverse nature of the syndrome. APAs are among the most common causes of a prolonged PTT in nonbleeding patients, as the APAs can bind phospholipids intended to trigger the coagulation cascade. This is the genesis of the "anticoagulant" part of the name "LA." This effect is utilized to test for the condition, as this effect can be overcome with the addition of larger amounts of phospholipids, similar to the substrate depletion effect in an immunoassay. A typical screening test is done either as an aPTT reaction with particularly low phospholipid concentration as to enhance the effect of the APA or through the use of dilute Russell viper venom, which is a phospholipid-sensitive venom that activates FX to FXa. In the presence of an APA, these assays will show a prolonged clotting time. This is typically followed by a confirmatory step that involves the addition of much larger quantities of phospholipids. In the case of a true APA, additional phospholipid will overwhelm the inhibitor and the clotting time will "correct." Current guidelines recommend that a ratio is determined by the lab, typically around 1.2, for the clotting time of the screen versus the confirm to call a positive result. Also, a lab may also choose to perform a mixing study to exclude the possibility of a factor deficiency, though the utility of this test varies depending on the specifics of the testing scheme. There are several other functional assays used for LA testing, such as the "silica clot time," which is a particularly LA-sensitive form of the aPTT assay, and aPTT assays using "hexagonal phase lipids," which are insensitive to LAs and thus can be used to "correct" the aPTT in the presence of an LA. Each of these methods has particular sensitivities and specificities for particular forms of LAs, and so each laboratory must determine the best fit for their patient population.

Another method is direct serological testing for antibodies against phospholipid-associated proteins. International guidelines recommend detection of either IgG or IgM antibodies to cardiolipin (aCL) or beta2 glycoprotein I (β2GP1). Elevation in IgA of either aCL or β2GP1 is of unknown clinical significance. APAs are identified by ELISA assays. The specimen from the patient is added to an ELISA well, and if the appropriate antibodies are present, they will bind to the target antigen. The anticardiolipin assay should be β2GP1-dependent. This is achieved by supplementing the reaction with either human or bovine β2GP1 that acts as a binding cofactor. The rationale behind this is that infections may induce transient aCL antibodies and these are typically not dependent on β2GP1. The laboratory definition of APS requires that a significant titer of aCL or β2GP1 persists for at least 12 weeks. International guidelines recommend that antibody titers should be >99th percentile of the population reference range.

While far more specific, a considerable number of APAs are not positive for either of these two proteins, and thus this testing is typically done in concert with the functional testing described above. APS is diagnosed symptomatically; thus the presence of an APA in a nonsymptomatic patient does not confirm a diagnosis. A diagnosis must include both the presence of an APA twice at least 12 weeks apart (as most APAs are transient and not clinically relevant), and accompanied by a patient history of at least one confirmed arterial thrombosis or VTE. It is important to note that several DOACs can cause false

positives on screening tests, most notably apixaban and rivaroxaban. Functional testing on patients currently taking these medications should be interpreted with extreme caution. If possible without high risk, patients should be asked to pause medication before testing to minimize interferences. Also, rivaroxaban has been shown to increase the risk of thrombosis in patients with an APA when compared to warfarin therapy, and any patient at risk for APS should not be treated with rivaroxaban until APS has been ruled out [17]. Heparin can also cause false prolongations on APA screening; however, most reagents include a heparin neutralizer to reduce this effect.

ADAMTS13 deficiency

Deficiency of the ADAMTS13 protein, responsible for cleaving ULVWFs, is linked to the development of a thrombotic microangiopathy termed "thrombotic thrombocytopenic purpura" (TTP), a life-threatening condition that can cause brain damage, stroke, and other thrombotic disorders, and can also lead to hemolytic anemia if untreated. Spurious activation of platelets through ULVWF causes the constant creation of thrombi, resulting in the sequestration and clearance of platelets leading to dangerous thrombocytopenia as well as vascular damage caused by the thrombi themselves.

The deficiency can be either congenital or acquired, though congenital deficiency of ADAMTS13 is very rare and is also known as Upshaw–Schulman syndrome or chronic relapsing TTP. This form of deficiency is only treatable through chronic plasma infusion or exchange from a healthy donor. Acquired ADAMTS13 deficiency is the result of an autoimmune inhibitor, and is also treated with plasmapheresis, though in most cases a sustained response can be achieved in several weeks.

While testing for ADAMTS13 is still relatively uncommon, it can be essential in the diagnosis and proper treatment of TTP. Currently, the standard of care states that therapeutic plasma exchange should be initiated as soon as TTP is seriously considered, given the rapid and destructive effects of TTP. However, testing is essential in confirming the diagnosis and should be performed on blood collected before exchange or infusion is initiated. Testing is functional; therefore both congenital and acquired forms can be tested in the same manner. Testing is typically either done using a fluorescence resonance energy transfer reaction using appropriately labeled peptides matching ADAMTS13's target sequence (between tyrosine 1605 and methionine 1606 of vWF), or through an ELISA assay with an immobilized target sequence and antibodies that recognize only the cleaved sequence.

Thrombophilia testing

A patient presenting with severe VTE, especially in repeated cases, will typically be worked up with a thrombophilia panel depending on the cause of the thrombotic event. If the VTE is a result of a provoked response (e.g., long plane ride and prolonged immobilization) additional workup is typically not performed. However, if the VTE was the result of an unprovoked response, then a thrombophilia workup is warranted. This panel should involve screening tests for APCR and PC and PS deficiencies. AT testing may also be performed, though deficiency should only be suspected in the case of high-dose heparin administration or if other conditions are ruled out.

Even though it is a prothrombotic condition, APS typically presents with a prolonged PTT and is the most common cause of prolonged aPTT without bleeding. A patient presenting with a prolonged PTT and VTE should be immediately tested for the presence of APAs if clinical suspicion is high for APS. When patients are undergoing a thrombophilia workup, anticoagulation therapy should be discontinued for at least 5–7 half-lives of the drug if possible without risk.

References

[1] P.A. Price, J.D. Fraser, G. Metz-Virca, Molecular cloning of matrix Gla protein: implications for substrate recognition by the vitamin K-dependent gamma-carboxylase, Proc. Natl. Acad. Sci. U.S.A. 84 (23) (1987) 8335–8339.

[2] D.K. Galanakis, J. Martinex, C. McDevitt, F. Miller, Human fetal fibrinogen: its characteristics of delayed fibrin formation, high sialic acid and AP peptide content are more marked in pre-term than in term samples, N. Y. Acad. Sci. Ann. 408 (1) (1983) 640–643.

[3] D.A. Quinn, R.B. Fogel, C.D. Smith, M. Laposata, B. Taylor Thompson, S.M. Johnson, et al., D-dimers in the diagnosis of pulmonary embolism, Am. J. Respir. Crit. Care Med. 159 (5 Pt 1) (1999) 1445–1449.

[4] H. Nomura, H. Wada, T. Mizuno, N. Katayama, Y. Abe, M. Noda, et al., Negative predictive value of D-dimer for diagnosis of venous thromboembolism, Int. J. Hematol. 87 (3) (2008) 250–255.

[5] P.S. Wells, Integrated strategies for the diagnosis of venous thromboembolism, J. Thromb. Haemost. 5 (Suppl. 1) (2007) 41–50.

[6] L. Shore-Lesserson, H.E. Manspeizer, M. DePerio, S. Francis, F. Vela-Cantos, M.A. Ergin, Thromboelastography-guided transfusion algorithm reduces transfusions in complex cardiac surgery, Anesth. Analg. 88 (2) (1999) 312–319.

[7] K. Görlinger, D. Dirkmann, A.A. Hanke, Potential value of transfusion protocols in cardiac surgery, Curr. Opin. Anaesthesiol. 26 (2) (2013) 230–243.

[8] K.P. Bharati, U.R. Prashanth, Von Willebrand disease: an overview, Indian. J. Pharm. Sci. 73 (1) (2011) 7–16.

[9] C. Ng, D.G. Motto, J. Di Paola, Diagnostic approach to von Willebrand disease, Blood 125 (13) (2015) 2029–2037.

[10] J.C. Gill, J. Endres-Brooks, P.J. Bauer, W.J. Marks Jr, R.R. Montgomery, The effect of ABO blood group on the diagnosis of von Willebrand disease, Blood 69 (6) (1987) 1691–1695.

[11] M.S. Enayat, A.M. Guilliatt, W. Lester, J.T. Wilde, M.D. Williams, F.G. Hill, Distinguishing between type 2B and pseudo-

von Willebrand disease and its clinical importance, Br. J. Haematol. 133 (6) (2006) 664–666.

[12] S. Sedano-Balbas, M. Lyons, B. Cleary, M. Murray, G. Gaffney, M. Maher, Acquired activated protein C resistance, thrombophilia and adverse pregnancy outcomes: a study performed in an irish cohort of pregnant women, J. Pregnancy 2011 (2011) 232840. Available from: https://doi.org/10.1155/2011/232840.

[13] R.A. Marlar, R.R. Montgomery, A.W. Broekmans, Diagnosis and treatment of homozygous protein C deficiency. Report of the Working Party on Homozygous Protein C Deficiency of the Subcommittee on Protein C and Protein S, International Committee on Thrombosis and Haemostasis, J. Pediatr. 114 (4 Pt 1) (1989) 528–534.

[14] A. Cuker, P.A. Gimotty, M.A. Crowther, T.E. Warkentin, Predictive value of the 4Ts scoring system for heparin-induced thrombocytopenia: a systematic review and meta-analysis, Blood 120 (20) (2012) 4160–4167.

[15] L.A. Linkins, A.L. Dans, L.K. Moores, R. Bona, B.L. Davidson, S. Schulman, et al., Treatment and prevention of heparin-induced thrombocytopenia: antithrombotic therapy and prevention of thrombosis, 9th ed: American College of Chest Physicians evidence-based clinical practice guidelines, Chest 141 (Suppl. 2) (2012) e495S–e530S. Available from: https://doi.org/10.1378/chest.11-2303.

[16] T.E. Warkentin, D.M. Arnold, I. Nazi, J.G. Kelton, The platelet serotonin-release assay, Am. J. Hematol. 90 (6) (2015) 564–572.

[17] V. Pengo, G. Denas, G. Zoppellaro, S.P. Jose, A. Hoxha, A. Ruffatti, et al., Rivaroxaban vs warfarin in high-risk patients with antiphospholipid syndrome, Blood 132 (13) (2018) 1365–1371.

Self-assessment questions

1. Choose the correct combination of sheer stress amount, operative platelet binding protein, and operative platelet receptor:
   a. high stress-vWF-GP Ia/IIa, low stress-collagen-GP Ib/IX/V
   b. low stress-vWF-GP Ib/IX/V, high stress-collagen-GP Ia/IIa
   c. high stress-vWF-GP Ib/IX/V, low stress-collagen-GP Ia/IIa
   d. low stress-vWF-GP Ia/IIa, high stress-collagen-GP Ib/IX/V
2. Which platelet feature contains small molecules such as ATP and serotonin?
   a. α granules
   b. Dense granules
   c. OCS
   d. Weibel−Palade bodies
3. Which group of factors is most directly affected by treatment by warfarin?
   a. Vitamin K-dependent factors
   b. Liver-dependent factors
   c. Protease-activated factors
   d. Propagation pathway factors
4. Which protein cross-links fibrin strands?
   a. FX
   b. FXI
   c. FXII
   d. FXIII
5. Which protein's action is potentiated by heparin?
   a. Antithrombin
   b. Protein C
   c. Protein S
   d. Protein Z-dependent protease inhibitor
6. Which laboratory test is used most to monitor warfarin therapy?
   a. aPTT
   b. PT
   c. D-dimer
   d. Fibrinogen
7. Which laboratory test is used most to monitor heparin therapy?
   a. aPTT
   b. PT
   c. D-dimer
   d. Fibrinogen
8. The D-dimer assay has a high negative predictive value for what patient presentation?
   a. High pretest probability of PE
   b. Low pretest probability of PE
   c. High pretest probability MI
   d. Low pretest probability MI
9. A patient presents to their PCP with the complaints of persistent heavy nosebleeds. Their PT and aPTT results are normal, and their provider is on the ball and orders vWF testing. Antigen levels are normal, and ristocetin activity is low. Factor VIII activity is normal. What type is the most likely given these results?
   a. Type 1
   b. Type 1C
   c. Type 2
   d. Type 3
10. What is the most common congenital factor deficiency?
    a. FVII
    b. FVIII
    c. FIX
    d. FX
11. Which platelet disorder results in platelets showing little to no aggregation with all agonists but ristocetin?
    a. Hermasky−Pudlak syndrome
    b. Bernard−Soulier syndrome
    c. Wiskott−Aldrich syndrome
    d. Glanzmann thrombasthenia
12. A patient presents to the ED with a DVT. A PT and aPTT are ordered, and they show an isolated increase in aPTT. What testing should be ordered next?
    a. Antiphospholipid antibody testing
    b. Platelet aggregometry
    c. D-dimer
    d. Serotonin release assay

Answers

1. C
2. B
3. A
4. D
5. A
6. B
7. A
8. B
9. C
10. B
11. D
12. A

Chapter 27

Diagnostic body fluid testing

Darci R. Block[1] **and Jonathan R. Genzen**[2]

[1]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, United States, [2]Department of Pathology, University of Utah School of Medicine & ARUP Laboratories, Salt Lake City, UT, United States

Learning Objectives

After reviewing this chapter, the reader should be able to:

1. State the regulatory requirements for body fluid validations according to CLIA.
2. Construct a differential diagnosis for pathological accumulations of common body fluids.
3. Cite the most clinically useful tests that aid in establishing or narrowing the differential diagnosis.
4. Interpret cellular, microbiological, and biochemical results for common body fluids examined in the clinical laboratory.

Introduction

Body fluids are routinely collected and sent to the clinical laboratory for analysis as a means to identify the pathological cause for an effusion or transudate. The measurement of biochemical markers in a particular fluid may give diagnostic insights into the pathology of the accumulation or underlying presence of disease. Often, the concentration of a body fluid analyte is compared with a reference interval, a predefined decision limit, or to the concentration of the analyte in blood, to facilitate result interpretation.

Laboratories must perform analytical validations and provide interpretive information for the body fluid tests they routinely perform. This requires a great deal of effort and resources that laboratories need to coordinate and prioritize. Situations where there is limited utility for measuring an analyte in a body fluid include scenarios where diagnostic information may be obtained from measurement of that analyte in blood (or another specimen source that is considered acceptable by the in vitro diagnostic manufacturer) or when the results do not impact patient management or care.

Regulations

The exact analytical validation requirements for a body fluid test may depend on local, state, and/or federal laws, as well as regulation from applicable accreditation agencies. In many jurisdictions, clinical laboratories that choose to test body fluid specimen types that are not specifically referenced in an assay manufacturer's instructions are required to analytically validate measurement in the alternative specimen source. For example, in the United States, the Clinical Laboratory Improvement Amendments of 1988 (CLIA'88) outlines federal regulations regarding establishment of performance specifications for modified Food and Drug Administration (FDA)-cleared assays (and/or assays not currently subject to FDA clearance or approval, e.g., laboratory developed tests), including accuracy, precision, analytical sensitivity, analytical specificity (interfering substances), reportable range, reference intervals, and any other performance characteristics required for test performance [1]. Similar requirements exist in other countries, including International Organization for Standardization (ISO) standards used widely in Europe and other parts of the world. The concerns that these regulations aim to address are due to the potential for matrix-related interferences that may be present in fluids that assay manufacturers do not have data for. Body fluids may have lower concentrations of protein, differences in pH, ionic strength, and viscosity, among other potential attributes that may contribute to an inaccurate analyte measurement. Several resources are now available to provide guidance for laboratories planning analytical validations of body fluid specimens [2,3].

Accuracy, termed more appropriately "trueness," is assessed through a variety of possible experiments to determine whether the test system in question is able to measure the true activity or concentration of a target analyte in a body fluid matrix despite an assay's original intended use in serum, plasma, or urine. There are multiple experiments that may be used to assess trueness in this context, including comparison to a reference method or the combination of calculating analyte recovery upon spiking solutions of standard concentration, mixing samples of high and low analyte concentration, and diluting

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00027-2

samples with an appropriate diluent [2–7]. Precision studies are performed as a means to assess the reproducibility of a method in a body fluid matrix. Experimentally, precision studies can be performed similarly to established approaches for evaluating precision in serum and plasma, choosing analyte concentrations that are near decision limits for the fluid of interest when possible. A detailed description of method validation is included in Chapter 4, Method validation, of this book.

Analytical sensitivity studies are performed to assess the reproducibility of a method at the lowest reportable concentration in a body fluid. Experimentally, sensitivity studies are similar to precision studies but are performed using low concentration specimens. Analytical specificity studies are performed to identify which substances cause interference, the extent of their effect on results, and to establish acceptable limits within which the lab can safely report accurate results. Typically, clinical laboratories adopt the manufacturer's recommended thresholds for interferences such as hemoglobin, lipids, and bilirubin without further study when verifying the performance of an assay on serum, plasma, or urine. Interference spiking studies are often needed when validating body fluid assays; however, there is potential for matrix effects with either the interferent itself or the presence of other substances not typically encountered in blood or urine specimens (e.g., meconium or very high glucose). Additionally, allowable bias and imprecision may be analyte- and fluid-dependent, as medical decision points in body fluids are often different as compared to those implemented for serum and plasma [3,8].

The reportable range represents the range of concentrations that an analyte can be accurately measured within, as well as the maximum reportable concentration with dilution. While reportable range studies in body fluids may be approached in a manner similar to serum and plasma, consideration should be given to choosing and validating appropriate diluents. Reference intervals allow the results of an analyte measurement to be interpreted either in the context of a normal reference population or for a defined population of subjects as a medical decision limit. Body fluids are typically only present in volumes large enough to aspirate for diagnostic purposes in the presence of pathological conditions, thus diminishing the utility of defining reference populations. Medical decision limits are commonly used; however, several challenges for establishing and/or verifying them include the lack of methodological detail from historical studies and the inability to predict the comparability of assays over time. Additionally, most studies do not provide evidence that their method is free of matrix bias [4]. These limitations confer significant difficulty to providing interpretive information or reference intervals with body fluid results. Studies that report the concentration of body fluid analyte in comparison to a corresponding serum or plasma specimen (or as a ratio) circumvent the latter concern. In most jurisdictions, laboratories are required to provide a mechanism to interpret the results for their tests; however, this may be viewed in a variety of formats including a printed/online resource or direct reporting into an appropriate information system.

Laboratories that are initiating or reevaluating their body fluid testing program should be cognizant of all the sources of variation with which they receive their specimens. The manner in which tests are ordered and the manner in which the fluid source is designated can impact the accuracy of information passed to the laboratory. Such workflows should be standardized as much as is practical. The laboratory should ensure that proper collection processes are followed, including use of acceptable collection containers and labeling that follows standard protocols for required content, including two patient identifiers, date and time of collection, and fluid source/anatomic site (if applicable). The preparation for and mode of transportation of specimens from the collection area, as well as the processing steps taken once received, should be conducted in accordance with established analyte stability limits.

Specific body fluids

Cerebrospinal fluid

Cerebrospinal fluid (CSF) is a plasma ultrafiltrate that is modified by both active and passive transports of substances from the systemic circulation into the central nervous system (CNS). CSF is produced primarily by the choroid plexus in the cerebral ventricles. Tight junctions between choroid plexus epithelial cells restrict and regulate the flow of water and solutes from underlying capillaries. The flow of CSF is directed from the lateral ventricles, through the interventricular foramina (foramen of Monro) to the third ventricle. CSF then flows through the cerebral aqueduct to the fourth ventricle, then through the lateral apertures (foramina of Luschka), the medial foramen (foramen of Magendie), or the central canal of the spinal cord. Through these pathways, CSF flows to the outer surfaces of the brain and spinal cord before exiting through arachnoid villi and granulations into dural venous sinuses or into extracranial lymphatics. CSF serves to cushion the brain and spinal cord mechanically during normal and traumatic movement. It serves an important role in nutrient delivery within the CNS as well as helping to remove metabolic waste products.

CSF analysis plays an important role in the diagnosis of many CNS disorders, including infections of the meninges (meningitis) and brain parenchyma (encephalitis). CNS infections can be due to bacteria, mycobacteria, viruses, fungi, or parasites. CSF analysis is also important in the identification of immune-mediated disorders of the CNS,

such as multiple sclerosis, as well as providing supportive evidence for prion disorders such as Creutzfeldt–Jakob Disease (CJD). CSF microscopic examination is also critical for the identification of malignant cells derived from (or metastatic to) the CNS. Along with radiologic investigation [primarily computerized tomography (CT) or magnetic resonance imaging (MRI)], CSF analysis is often conducted in patients who have unexplained mental status changes, fever, headache, meningeal symptoms, and/or suspected cranial nerve abnormalities on physical examination.

Most CSF specimens are obtained through a procedure called lumbar puncture (LP; or spinal tap). In this procedure, patients are placed in a lateral decubitus position while a needle is placed into the subarachnoid space in the lumbar region of the spinal canal. Some LPs are conducted using radiologic and fluoroscopic guidance. CSF specimens (particularly in neurologic intensive care settings) may also be obtained through preexisting ventricular shunts or drains. When the LP procedure causes bleeding into the subarachnoid space, this is called a traumatic tap. Differentiation of traumatic tap from subarachnoid hemorrhage can usually be made based on laboratory findings as discussed later in this section.

Upon successful LP, CSF opening pressure is measured with a manometer. Normal CSF opening pressure in adults is approximately 50–180 mm H_2O, with pressures being lower in young children and slightly higher in obese patients. Abnormal or obstructed CSF flow can lead to accumulation of CSF in the brain (hydrocephalus) and increased intracranial pressure. In patients with elevated CSF opening pressure ($>$200 mm H_2O), only 1–2 mL of CSF should be removed for diagnostic purposes, as collection of additional volume carries an increased risk of potentially fatal cerebral herniation. In these cases, medical intervention to control intracranial pressure is warranted.

After measurement and documentation of opening pressure, CSF is typically collected into three or four consecutively numbered tubes (e.g., tubes 1–4). There are several reasons for collection into multiple tubes. Given the invasiveness of CSF collection, it is imperative that adequate specimen is collected for subsequent diagnostic testing, which may be directed to several different laboratory sections such as chemistry, hematology, microbiology, molecular, and cytology. While laboratories have varying protocols, tubes 1 or 2 are usually reserved for routine chemistry, immunology, and/or esoteric testing, as these assays are less likely to be negatively influenced by traumatic taps. The next tube is often reserved for microbiology (Gram stain, culture, and/or infectious disease molecular testing), while the last tube(s) are often used for hematology, cytology, and/or additional molecular or esoteric testing. Performing cell counts on the final tube decreases the influence of traumatic taps on the CSF white blood cell (WBC) count. Many laboratories will also perform an additional cell count on the first tube, as differences in cell counts between tubes 1 and 4 are indicative of a traumatic tap. In suspected traumatic taps, correction equations can be used to adjust CSF WBC counts for the influence of peripheral blood contamination. Finally, laboratory distribution of CSF should be modified based on a patient's unique clinical circumstances. For example, if the primary purpose of the LP is to evaluate for the possibility of CNS malignancy, specimen volume priority should be directed to hematology and cytology as opposed to chemistry and microbiology.

Normal CSF is clear and relatively colorless in appearance. Yellow or pink coloration is called xanthochromia, which is often indicative of prior bleeding (e.g., subarachnoid hemorrhage) into the CSF. In such cases, red blood cell (RBC) destruction leads to hemoglobin breakdown and accumulation of bilirubin. Yellow coloration, however, may also be evident with systemic hyperbilirubinemia or increased CSF total protein.

CSF cell counts (RBC, WBC, and differential) should be conducted shortly after collection to prevent falsely low results due to cell lysis and/or settling. Cytospin preparations are often used for the differential to improve visualization and identification of cells, given the relatively low cell counts in CSF versus serum. Normal CSF contains $<$5 RBC/μL in both children and adults. Normal adult CSF has $<$5 WBC/μL, while neonatal CSF typically has $<$30 WBC/μL, with progressively decreasing numbers through childhood. These ranges have been based on historical evidence from manual cell counts, typically conducted using a hemocytometer. Recent trends toward automating body fluid cell counting methods, primarily flow cytometry, are forcing many laboratories to evaluate low-end bias and imprecision near assay cutoffs. Establishing method-specific CSF cell count reference ranges using these techniques may challenge the notion of the current, traditionally accepted ranges that permeate the literature and clinical practice.

The WBC differential in normal adult CSF typically shows a lymphocyte predominance (30%–90%), with a moderate percentage of monocytes (15%–50%), and a small number of PMNs (0%–6%). Eosinophils and ependymal cells may be observed, albeit less frequently. The increased WBC count in bacterial meningitis is usually accompanied by a predominance of PMNs, while viral meningitis often shows a lymphocyte predominance. The WBC differential is not specific for bacterial, viral, and fungal meningitis, however, as cell-type predominance can vary between patients and time course of infection. Cytology may be essential when abnormal cells are observed in the differential or when malignancy is suspected.

Normal CSF total protein varies based on age, with adult concentrations typically in the range of 15–45 mg/dL. Neonatal total protein concentrations are typically higher, in the range of 40–120 mg/dL and are higher in

preterm infants. CSF total protein concentrations can increase in the context of infection, inflammation, malignancy, and/or damage to the blood−CSF barrier. Elevations of CSF total protein are particularly common in bacterial and or fungal infections. Normal CSF glucose concentrations are approximately 2/3 the concentration of a corresponding serum specimen, and a normal glucose ratio is often expressed as 0.6 in the literature. CSF reference intervals are also often presented as a range from approximately 50−80 mg/dL. Decreased CSF glucose is frequently observed in bacterial and fungal CNS infections. CSF glucose concentrations may be relatively normal in viral CNS infections.

Gram stain and culture remain essential for the routine diagnosis of bacterial meningitis. The majority of neonatal cases of bacterial meningitis are due to group B *Streptococcus*, *Escherichia coli*, and more rarely *Listeria monocytogenes*. From infancy to young adulthood, the prevalence shifts toward *Neisseria meningitidis*, *Streptococcus pneumoniae*, and *Haemophilus influenzae*. Vaccination against *H. influenzae* type B, however, has decreased the incidence of meningitis caused by this organism. A diagnosis of neurosyphilis (tertiary infection due to the fastidious bacteria *Treponema pallidum*) is supported by a positive CSF Venereal Disease Research Laboratory (VDRL) test result.

Significant advances in molecular techniques have made nucleic acid-based testing (NAT; e.g., PCR) fundamental to the rapid detection of many viral CNS infections. NAT is now the mainstay for diagnosis of commonly observed CNS viral infections such as enteroviruses and members of the herpesviridae family (e.g., HSV-1, HSV-2, CMV, and EBV). Mycobacterial infections may require a combination of culture and/or molecular testing for their detection. Fungal infections may be detected though a combination of culture, direct smears, special preps (e.g., India ink for *Cryptococcus neoformans*), antigen testing, and/or molecular methods.

Testing for CNS production of immunoglobulins plays an important role in the assessment of inflammatory disorders such as multiple sclerosis. CNS inflammatory disorders may be accompanied by the production of immunoglobulins within the brain, spinal cord, or intrathecal space (under the arachnoid membrane). Qualitative assessment of immunoglobulin production with isoelectric focusing, high-resolution gel electrophoresis, and/or immunofixation electrophoresis of paired CSF and serum specimens allows for the detection of oligoclonal bands that are present in the CSF but absent in serum, a finding that is often observed in multiple sclerosis and other CNS inflammatory processes (see Fig. 27.1).

Several calculations can also be used to evaluate the integrity of the blood−CSF barrier as well as intrathecal synthesis of immunoglobulins. Changes in blood−CSF barrier permeability can be assessed through calculation of the Albumin Index (note units):

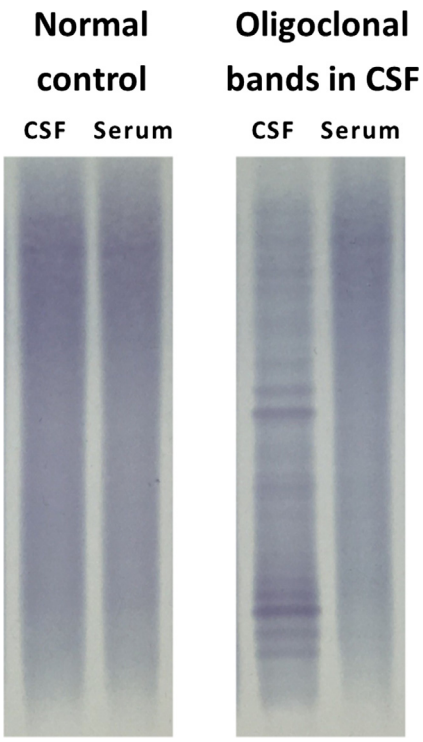


FIGURE 27.1 Cerebrospinal fluid oligoclonal band analysis. Paired cerebrospinal fluid and serum specimens evaluated for the presence of oligoclonal bands using an isoelectric focusing method combined with immunofixation. *Left*: no evidence of oligoclonal bands in cerebrospinal fluid or serum from this normal control. *Right*: oligoclonal bands observed in cerebrospinal fluid, with no matching bands observed in the patient's corresponding serum.

$$\text{Albumin index} = \frac{\text{CSF albumin}\left(\text{mg/dL}\right)}{\text{Serum albumin}\left(\text{mg/dL}\right)} \times 1000$$

Albumin Index results <9 are consistent with an intact blood−CSF barrier, while results >30 are consistent with severe barrier impairment. Intermediate results indicate mild-to-moderate impairment of the blood−CSF barrier. CSF IgG synthesis can be assessed through calculation of the IgG Index (note units):

$$\text{IgG index} = \frac{\text{CSF IgG}\left(\text{mg/dL}\right)}{\text{CSF albumin}\left(\text{mg/dL}\right)} \times \frac{\text{Serum albumin}\left(\text{mg/dL}\right)}{\text{Serum IgG}\left(\text{mg/dL}\right)}$$

IgG Index results <0.7 are considered to be normal. It should be noted that any cause of CSF immunoglobulin synthesis (e.g., inflammatory or infectious) can lead to increased IgG Index results; therefore elevated values are not specific for multiple sclerosis. Additionally, impairment in the blood−CSF barrier can lead to increased CSF IgG and albumin (from the peripheral circulation), and this can also impact the interpretation of IgG Index results. Some

laboratories also calculate CSF IgG Synthesis Rate (note units):

$$\text{IgG synthesis rate} = \left[\left(\text{CSF IgG} - \frac{\text{Serum IgG}}{369}\right) - \left[\left(\text{CSF albumin} - \frac{\text{Serum albumin}}{230}\right) \times \left(0.43 \times \frac{\text{Serum IgG}}{\text{Serum albumin}}\right)\right]\right] \times 5\ \text{dL/day}$$

Units for all proteins in the IgG Synthesis Rate equation are expressed as mg/dL. The equation includes several constants, which are estimates derived from normal populations. For example, "369" is an approximate normal serum:CSF ratio for IgG, "230" is an approximate normal serum:CSF ratio for albumin, "0.43" is a ratio of the molecular weight of albumin to the molecular weight of IgG, and "5" is an estimate of the average daily CSF production in adults (expressed in dL) [9,10].

Identification of CSF leakage is important in the context of nasal drainage (rhinorrhea) or ear drainage (otorrhea) that may be associated with trauma, cranial surgery, or neoplasm. β2-Transferrin is often used as a marker for CSF in these circumstances, as neuraminidase in the CNS removes sialic acid from transferrin to form the desialated (β2) form. Paired testing with a serum specimen, however, is usually performed, as this helps to eliminate the possibility that a patient may have a disorder in which serum transferrin is also hypoglycosylated.

Transmissible spongiform encephalopathies are rare, progressive fatal neurodegenerative disorders caused by abnormally folded proteins known as prions. Prion disorders in humans include CJD, kuru, fatal familial insomnia, and Gerstmann–Sträussler–Scheinker syndrome [11]. These prion diseases involve conformational changes to the protein product of the *PNRP* gene (PrP) [12]. PrP is a cell surface glycoprotein which is present normally in an α-helix-rich conformational state (PrP^C). In prion diseases, PrP^C undergoes a conformational change to form β-sheet-rich isomers known as PrP^{Sc}. PrP^{Sc} is less soluble and has a propensity to form amyloid fibrils. Initiation of conformational changes in PrP^C can occur sporadically, it can be triggered by exposure to PrP^{Sc} from exogenous sources, or alternatively mutations in the *PNRP* gene can also result in a propensity for PrP^C misfolding. While most cases of CJD are sporadic, a minority of cases are associated with inherited mutations. Variant CJD (vCJD) is the human prion disease associated with consumption of meat from cows affected by bovine spongiform encephalopathy ("mad cow disease").

Diagnosis of prion diseases (such as sporadic and variant CJD) is typically supported by clinical, radiologic (MRI), and encephalographic evidence. Biopsy may be considered in certain cases. CSF studies may be supportive and include analytes such as protein 14-3-3, tau, and neuron-specific enolase. Advancements in prion diagnostics include real-time quaking-induced conversion assays for the detection of PrP in CSF [13]. Molecular testing of blood specimens may also be used to detect inherited *PRNP* mutations. Laboratory testing for prion diseases is performed primarily by referral laboratories and prion surveillance centers. This is for both practical logistical reasons (due to the rarity of prion disorders) and the safe handling of potentially infectious tissue specimens. Appropriate handling and disposal are key considerations in laboratory analysis of specimens from patients with suspected prion disease [14]. The World Health Organization has published infection control guidelines that describe the level of infectivity and risk across a variety of tissues and specimen types [15].

Pleural fluid

A small amount of pleural fluid is normally present between the parietal and visceral membranes of the thoracic cavity surrounding the lungs. The parietal membranes line the inner surface of the chest wall, while the visceral membranes line the lungs. Pleural fluid serves as a lubricant between the lungs and inner chest wall during breathing. It is produced under normal conditions by capillary hydrostatic and osmotic pressure from the parietal vasculature. Pleural fluid production is balanced by absorption, predominantly by lymphatic stomata (openings into lymphatic capillaries on mesothelial surfaces) and venous uptake.

Increases in pleural fluid production, as well as decreases in lymphatic absorption, can lead to fluid accumulations known as pleural effusions. These effusions may be unilateral or bilateral depending on their etiology. Some causes for pleural effusions include increased pressure in the lung circulation (e.g., with congestive heart failure or pulmonary emboli), changes in pulmonary vascular permeability that may accompany inflammation or infection, accumulation of pus during infection (empyema), movement of nonpleural fluids (e.g., peritoneal fluid or lymphatic drainage) into the thoracic cavity, hemorrhage, or a variety of disorders, which slow or block the absorption of fluid back into the lymphatic capillaries [16].

Many small pleural effusions may be clinically asymptomatic and detected only during radiologic studies such as chest X-ray, CT, MRI, or ultrasound (US). Pleural effusions may also be suspected based on clinical symptoms or physical examination of the chest using a stethoscope. Patients with symptomatic pleural effusions may report shortness of breath, difficult or painful breathing, or general pulmonary symptoms such as cough [17]. They may also be febrile if an underlying infection is present.

Pleural fluid can be collected through a procedure known as thoracentesis. This may be performed diagnostically (to obtain fluid for cellular, chemical, and/or

microbial analyses) or therapeutically to remove fluid accumulations when clinically indicated. Not all pleural effusions require diagnostic or therapeutic thoracentesis. For example, a bilateral pleural effusion in the context of known congestive heart failure may resolve after initiation of diuretic therapy. Thoracentesis is often conducted using US guidance of the collection needle, as the use of US increases the likelihood of a successful collection and decreases the risk of serious complications such as pneumothorax (collection of air in the pleural cavity) and hemothorax (collection of blood in the pleural cavity) [18]. Immediately after collection, the pleural fluid specimen can be transferred into several different containers to enhance diagnostic testing. For example, transfer of fluid into an ethylenediaminetetraacetic acid (EDTA)-containing blood collection tube can decrease the chance of clotting before cellular analysis. Transfer of fluid into blood culture bottles has been shown to increase the rate of bacterial identification. Additionally, transfer of fluid into a capped, heparinized syringe improves the accuracy of pH determination in the matrix [19]. Care should be taken to avoid cross contamination with collection and transport container additives that may be incompatible with subsequent clinical laboratory testing. Contamination during processing and handling should also be avoided with specimens intended for microbiological and/or molecular analyses.

The most important first step in pleural fluid analysis is to distinguish exudates from transudates. Transudates are fluid accumulations that are caused by differences in hydrostatic and osmotic pressures across otherwise healthy membranes. Exudates are fluid accumulations that may be due to membrane damage, infection, inflammation, or malignancy. The gross appearance of a pleural fluid specimen (e.g., clear, translucent, turbid, bloody, lipemic, or purulent) may provide some initial clues in this differentiation. For example, transudates are often (but not always) relatively clear or translucent. Gross appearance, however, is not diagnostic, and more effective strategies have been developed.

Differentiation of pleural fluid exudates from transudates is usually conducted using Light's criteria or associated parameters (Table 27.1) [24]. Pleural fluid total protein, lactate dehydrogenase (LDH), and albumin—along with corresponding serum measurements—can help to differentiate exudates from transudates. A pleural fluid is typically considered to be an exudate if it has a fluid-to-serum protein ratio of >0.5, a fluid-to-serum LDH ratio of >0.6, or a fluid LDH activity that is >2/3 the upper limit of a normal serum LDH activity [17,24]. Heart failure and diuretic-associated misclassifications by Light's criteria (scenarios where transudates are falsely categorized as exudates) may be correctly categorized as

TABLE 27.1 Differentiation of pleural effusion exudates and transudates.

	Transudate	Exudate
Appearance		
Clarity	Clear	Opaque, lipemic, purulent, and bloody
Color	Pale yellow; straw	Variable
Propensity to clot	Rarely	Sometimes
Light's criteria		
Total protein ratio (fluid/serum)	≤0.5	>0.5
LDH ratio (fluid/serum)	≤0.6	>0.6
LDH activity	≤2/3 upper limit nl serum	>2/3 upper limit nl serum
Prevention of misclassification with heart failure or diuretic use		
Albumin gradient (serum-fluid)	>1.2 g/dL	≤1.2 g/dL
Protein gradient (serum-fluid)	>3.1 g/dL	≤3.1 g/dL
Additional criteria		
Bilirubin ratio (fluid/serum)	≤0.6	>0.6
Cholesterol	≤45 mg/dL	>45 mg/dL
Cholesterol ratio (fluid/serum)	<0.3	≥0.3

Notes: Refer to [16,17], and [19–23].
LDH, Lactate dehydrogenase.

transudates if a serum-to-fluid albumin gradient (serum albumin minus fluid albumin) is >1.2 g/dL or a serum-to-fluid protein gradient (serum protein minus fluid protein) is >3.1 g/dL [20]. Exudates often have a pleural fluid-to-serum bilirubin ratio >0.6, a pleural fluid cholesterol >45 mg/dL, or a pleural fluid-to-serum cholesterol ratio >0.3, although these tests do not typically add diagnostic value significantly beyond Light's criteria and are not recommended for routine testing.

Transudative pleural effusions are most frequently associated with congestive heart failure, where circulatory overload leads to increased osmotic pressure in the lung capillaries. Other causes of transudative effusions include decreased serum albumin (e.g., cirrhosis, nephrotic syndrome, or protein-losing enteropathy), hepatic hydrothorax (where peritoneal fluid crosses the diaphragm with portal hypertension or leak), peritoneal dialysis (PD), collapsed lung (atelectasis), constrictive pericarditis, or rarely urinothorax due to obstruction or trauma [16,17]. Exudative effusions may be caused by a number of mechanisms, some of which include malignancy, pulmonary embolism, parapneumonic effusions (pleural effusions which occur with pneumonia), tuberculosis, rheumatologic and inflammatory disorders, pancreatitis, or chylothorax [16,17].

Gram stain and culture are essential for the assessment of potential bacterial or fungal infection. Cytology should be routinely ordered on pleural fluid specimens to evaluate for potential malignancy. Cell counts, including RBC and WBC counts with WBC differentials can provide supportive evidence for a variety of potential etiologies, in particular lymphocyte and neutrophil predominant effusions. While low pH (<7.2) is not specific for pleural fluid infection, it is one criteria used by clinicians in deciding whether tube drainage will be required for therapeutic management of pleural effusions associated with infection [25]. pH analysis of pleural fluids is typically conducted using blood gas analyzers to provide more accurate and rapid results. Some blood gas analyzers are now FDA-cleared for this particular specimen type. When tuberculosis is suspected, elevated adenosine deaminase (ADA) may also provide supportive evidence of mycobacterial infection. Routine use of tumor markers in pleural fluid specimens is not recommended. Some studies have shown that tumor marker testing (such as CA19-9 and CEA), along with corresponding serum levels, may provide supportive information that could be used along with cytology and other clinical evidence to support the diagnosis of malignant effusion [26]. Current pleural disease guidelines do not support the routine analysis of tumor markers in pleural effusions [27].

Peritoneal fluid/ascites

The peritoneal cavity is the space contained between the peritoneal membranes surrounding the abdomen. A small volume of serous fluid, 25–50 mL, is normally present within this space, lubricating the surfaces and maintaining pliability [28]. The mechanisms of normal fluid filtration out of systemic microvessels follow Starling's equation, which includes liquid conductance (k), vascular permeability of the peritoneal capillaries (s), hydrostatic pressure gradients (**P**) across sinusoidal (ss) and hepatic interstitial (hi) compartments, and colloid oncotic pressure differences (π) attributable primarily to serum albumin concentrations [29,30].

$$\text{Flow} = k\left[(P_{ss} - P_{hi}) - s(\pi_{mv} - \pi_{if})\right]$$

When these forces overwhelm the absorptive capacity within the peritoneum, primarily through drainage via the lymphatic system (maximally 850 mL/day), fluid accumulates [28]. The pathologic accumulation of fluid within the peritoneal cavity is commonly referred to as ascites. The most common cause of ascites is liver cirrhosis resulting from alcohol abuse, viral hepatitis, or tuberculous infection, and causation frequency is location and population dependent. Ascites can be diagnosed by physical examination noting bulging flanks or shifting dullness to percussion and may be aided by abdominal imaging if the volume is relatively small (<500 mL) or the patient is obese.

Paracentesis is a percutaneous procedure whereby a 1- to 4-in. needle is inserted traditionally into the midline between the pubis and umbilicus, to aspirate fluid under US guidance to avoid puncturing major vessels, organs, or the bowel. Today, the lower left quadrant is the preferred location as the prevalence of abdominal obesity increases and more large volume fluid reductions are performed. Up to 60 mL of fluid may be removed for diagnostic purposes in patients presenting with new onset ascites or to evaluate new signs and symptoms concerning for infection in those patients with existing ascites [28]. Alternatively, a larger bore needle may be inserted for therapeutic purposes to drain larger volumes of fluid (>5 L) to relieve symptoms of tense ascites, which may include a grossly distended abdomen, shortness of breath, and early satiety (feeling full after a small meal). There are few reported contraindications to performing paracentesis, even in cirrhotic patients with significant coagulopathies (interquartile range of PT-INR 1.4–2.2) and thrombocytopenia (interquartile range of platelets = 42–56 × 10^3/L) [31]. Patients with clinically evident coagulopathies such as disseminated intravascular coagulation should have paracentesis deferred [32].

The diagnostic approach to peritoneal fluid analysis in patients presenting with ascites involves characterizing the fluid as sterile (i.e., noninfectious), infectious, and/or malignant. The most common cause of ascites is portal hypertension related to cirrhosis, representing approximately 80%

of cases throughout North America and Europe [33]. It is important to consider noncirrhotic causes of ascites in the differential. These include prehepatic etiologies, such as malignancy, pancreatitis, portal vein thrombosis, lymphatic injury or obstruction, bowel perforation, renal failure, and tuberculosis [28]. Posthepatic etiologies include congestive heart failure, constrictive pericarditis, Budd–Chiari syndrome, and veno-occlusive disease [28]. The gross appearance of peritoneal fluid should be noted as it may have diagnostic clues. Ascites due to liver cirrhosis typically has a clear straw-colored appearance. A bloody appearance may indicate traumatic tap (often clots upon standing), malignancy, tuberculosis, pancreatitis, or abdominal trauma [33]. A turbid appearance may indicate the presence of chylomicrons (chylous ascites; which is sometimes associated with malignancy) or neutrophils (pseudochylous ascites) associated with infection.

The recommended screening tests for peritoneal fluids collected for diagnostic purposes received by the clinical laboratory include measurement of a cell count with differential, culture, albumin, and total protein. The serum ascites albumin gradient (SAAG) is the calculated difference between the concentration of albumin in the ascitic fluid and serum, ideally collected within 30 minutes up to 24 hours of each other [34]. SAAG has been shown to correlate directly with measures of portal pressure, and a cutoff of ≥ 1.1 g/dL identifies portal hypertension as the cause of ascites with 97% sensitivity [34]. Cirrhosis is the most common cause of increased portal hypertension but may also be caused by portal vein thrombosis and the posthepatic causes listed above [35].

Spontaneous bacterial peritonitis (SBP) is a frequent and significant complication of cirrhosis affecting 10%–30% of patients with ascites [33]. This infection occurs when intestinal bacteria migrate into the peritoneal cavity as a consequence of the compromised integrity of the GI tract, the type and burden of bacteria present, and any potential reduced immune function of the host. Symptoms vary widely from being completely asymptomatic, presenting with abdominal pain, and in extreme cases accompanied by hepatic encephalopathy.

The primary finding upon cellular analysis of ascitic fluid is elevated polymorphonuclear (PMN) leukocytes or neutrophil count. The recommended cutoff to provisionally diagnose infection and initiate antibiotic therapy is >250 cells/μL [28]. Cultures should be collected for all cases of suspected infection into a culture bottle at the bedside, as this has been demonstrated to maximize the sensitivity of recovering an organism [36]. Leukocyte esterase measured by point of care reagent strips has been studied as a surrogate for cellular analysis for screening purposes or where cellular analysis is not readily available. One diagnostic dilemma encountered in these cases is differentiating spontaneous from secondary bacterial peritonitis. Secondary causes include those that can be surgically corrected such as bowel perforation or loculated abscess in the absence of perforation. The diagnosis of SBP is considered when secondary causes have been ruled out. Runyon observed 100% sensitivity for identifying secondary bacterial peritonitis due to perforation in neutrocytic ascites with or without liver disease when total protein is >1.0 g/dL, LDH is > upper limit of normal for serum, and glucose is <50 mg/dL [28,37]. Nonperforation secondary peritonitis is evident when the PMN count increases from baseline upon repeat paracentesis performed 48 hours after initiating antibiotic therapy [36].

The differentiation of cardiac ascites from cirrhotic ascites is a common clinical conundrum as they are the two most common causes of ascites presenting with elevated SAAG. Heart failure leads to the development of high gradient ascites due to hepatic sinusoidal hypertension. Since the sinusoids are normal and have not been damaged from collagen deposition associated with cirrhosis, protein tends to "leak" more readily into ascites and is associated with higher total protein concentrations with a cutoff >2.5 g/dL used to distinguish it from cirrhotic ascites [33]. Misclassification does occur in milder cases of cirrhotic ascites where synthetic output of the liver is not as significant and in heart failure patients with hypoalbuminemia. It has, therefore, been suggested in limited studies to use serum heart failure biomarkers such as B-type natriuretic peptides in addition to ascitic fluid analysis to screen patients presenting with new onset ascites that could identify heart failure as the cause rather than rely solely on the pattern of elevated SAAG and ascitic fluid total protein [38]. Refer to Chapter 30, Biomarkers of Coronary Artery Disease and Heart Failure, for a complete description of lab testing for heart failure.

SAAG < 1.1 g/dL is not consistent with the presence of portal hypertension and occurs commonly in patients with malignancy-related ascites, tuberculosis infection, nephrotic syndrome, and pancreatitis. Malignancy-related ascites develops as a result of multiple mechanisms including metastatic infiltration of tumor cells through the peritoneum (peritoneal carcinomatosis), liver metastasis, portal or hepatic vein compression, or obstruction within the lymphatic system leading to chylous effusion [39]. Less commonly, the malignant ascites is caused by a primary peritoneal malignancy. More commonly, it is caused by secondary peritoneal carcinomatosis from breast, ovary, endometrium, GI tract, or pancreatic cancers. Hepatocellular carcinoma is a common malignancy of the liver often associated with cirrhosis caused primarily by chronic hepatitis.

The diagnosis of malignant ascites relies on the identification of malignant cells from either cytologic analysis of fluid or tissue biopsy. Cytology suffers from relatively poor sensitivity, as recovery of cells from fluid is often

limited. Therefore biomarkers that can distinguish malignant from nonmalignant ascites are helpful in order to reduce the number of unnecessary biopsies and aid in diagnosis. Numerous tumor markers have been measured in ascitic fluids, including carcinoembryonic antigen (CEA), α-fetoprotein (AFP), cancer antigen (CA) 19-9, CA 125, and CA 15-3, among others. Most studies report decision limits that are derived from differentiating a population with malignant ascites from a population with nonmalignant ascites. These studies report widely varying sensitivities and specificities dependent on which malignancies were included in the study population and whether they are known secretors of that protein, as demonstrated by measurement in serum [40]. The most promising tumor marker for its positive predictive value is CEA, with a pooled sensitivity of 43% (95% CI [38%, 48%]) and specificity 96% (95% CI [93%, 97%]) reported from a meta-analysis of seven studies with cutoffs ranging from 3.5 to 25 ng/mL [41]. These decision limits are comparable to those reported for serum, and just as with any immunoassay, results and decision limits should not be compared between manufacturers due to the lack of harmonization. Elevated cholesterol has been shown to identify malignant ascites with sensitivities and specificities >80% using varying cutoffs between 32 and 70 mg/dL, owing to increased vascular permeability, release from tumor cells, and lymphatic obstruction [4]. The measurement of cholesterol and other tumor markers is probably best limited to circumstances when malignancy is suspected and cytology is negative or indeterminate for best results [33].

Chylous ascites is a rare disorder caused by disruption of the lymphatic system wherein chyle leaks into the abdominal cavity causing a milky-appearing, triglyceride-rich effusion in the peritoneum. The physical appearance can be deceiving as elevated white cell counts may also give a turbid appearance. The most common etiology of chylous ascites in adults is malignancy and cirrhosis, while in children it is congenital lymphatic disorders [42]. Infectious causes such as tuberculosis causing lymphatic obstruction or disruption are more prevalent in developing countries. Triglycerides may be measured to confirm the existence of chylous ascites with a recommended cutoff of >200 mg/dL [43]. Notably, cholesterol measurement does not differentiate chylous from nonchylous effusions and should not be tested for the purpose [44].

Tuberculous ascites is observed in countries with increased prevalence of infection. Acid-fast bacilli (AFB) smears and cultures are used to identify mycobacterial infections. The sensitivity of the smear, however, is quite limited and culture results may take up to several weeks to obtain. The American Association for the Study of Liver Diseases guidelines recommend that only patients at high risk of tuberculous peritonitis (e.g., recent immigration from an endemic area or coinfected with an acquired immunodeficiency syndrome) should be tested, and in such cases, biopsy with mycobacterial culture of tubercles is most efficient. Many studies have been conducted measuring ADA in ascitic fluid with a cutoff >40 U/L giving a pooled sensitivity of 93% (95% CI [89%, 95%]) and specificity 96% (95% CI [94%, 97%]) in a metaanalysis of 16 studies utilizing the Giusti method [45]. The Giusti method involves the colorimetric quantification of ammonia as a colored phenol end product (Berthelot's reaction) from enzymatic conversion of adenosine (substrate) by ADA [46]. This approach is recommended for facilities that may not have biopsy services available or for screening patients to identify those that would benefit most from the procedure.

The measurement of peritoneal glucose is not routinely recommended but may be useful as an adjunct in cases of suspected peritoneal infection. Low glucose concentration is a nonspecific finding, as any process where the localized metabolic activity is increased may result in decreased glucose concentrations. It is recommended to compare the fluid glucose concentration to serum or plasma concentrations for interpretation.

Pancreatic enzyme concentrations two- to threefold greater than corresponding serum concentrations may be highly suggestive of pancreatic ascites though not entirely specific [33]. Measurement of amylase and/or lipase in abdominal fluid may aid in the identification of pancreatic fistulas caused by chronic pancreatitis or formation of a pancreatic fistula following surgery. However, in the setting of a low gradient SAAG and elevated serum enzymes, quantitation of abdominal fluid amylase and/or lipase is unlikely to alter patient management.

Creatinine and/or urea may be elevated in ascitic fluid that is contaminated with urine. Fluid to serum ratios >1 for either creatinine or urea are highly suggestive of bladder rupture or perforation following surgery and trauma or the development of urinary fistula. Such findings will elicit a more focused radiologic examination and surgical intervention to repair.

Choleperitoneum is defined as the presence of bile in the peritoneum, which is an infrequent yet highly treatable cause of ascites observed most commonly as a complication following laparoscopic cholecystectomy or following other surgeries or trauma affecting the biliary tree [47]. Nuclear medicine cholescintigraphy scans are used to diagnose damage to the gallbladder and biliary system but are expensive and time consuming. A fluid that is brown or green highly suggests the presence of bile; however, some bilious fluid has been described as yellow, amber, and orange, and can therefore be easily mistaken as normal. Measurement of the fluid total bilirubin to serum ratio with cutoff >5 has 100% diagnostic accuracy for detecting bile. Detection of bile leaks using the fluid to serum bilirubin ratio following orthotopic

liver transplant is reportedly less accurate owing to possible dilution within the peritoneal cavity due to preexisting ascites, with an optimal cutoff of >3.25 (area under the curve: 0.89, sensitivity 73%, specificity 95%) [48]

Ascitic fluid methods

The accuracy for the measurement of numerous analytes using both chemistry and immunoassay methods have been confirmed in ascitic fluid using the Roche cobas 8000 platform [3,5] and for tumor markers on the Beckman Coulter UniCel DxI 800 platform [7,40]. Older studies have suggested that colorimetric albumin methods should be avoided when measuring SAAG as they are less specific for albumin (demonstrating bias compared to immunochemical methods) with lack of analytical sensitivity in the measurement of low concentrations of albumin [33,49]. Bias has been reported between bromocresol green and purple (BCP) methods in serum owing to BCP having improved specificity for human albumin. This bias was not observed, however, in a small cohort of peritoneal fluids, which caused several SAAG values to be reclassified from high to low gradient when using the BCP method [50]. These findings need to be replicated in a larger cohort; however, they do elucidate a concern laboratories should be aware of.

The gold standard for performing body fluid cell counting is use of a hemocytometer equipped with a manual counting chamber. More recently, automated cell counting analyzers have become available with manufacturer's intended use including quantitation and differentiation of cells in body fluids. Manual methods cannot be completely replaced, however, as automated methods may have limited analytical sensitivity. Use of automated methods may also require manual microscopy to identify abnormal cells flagged by the instrument [51].

The use of leukocyte esterase reagent strips marketed for dipstick urinalysis has been studied as a rapid and simple point of care test for surrogate cellular analysis to expedite identification and treatment of SBP [33]. While studies demonstrate significant correlation to neutrophil count, diagnostic accuracy varies as the reading is semiquantitative, have subjective interpretation, and the cutoffs considered positive need to be established as they vary from manufacturer to manufacturer [52]. It has been suggested that reagent strip methods may be suitable for screening in places where manual or automated cell counting methods are unavailable, such as in developing countries [52].

Dialysate

PD is a type of ambulatory dialysis that can be employed in the home for patients with end-stage kidney disease. The procedure involves infusing 2 L of a hyperosmotic fluid through an abdominal catheter into the peritoneal cavity where the peritoneum acts as an endogenous dialysis membrane to remove metabolic waste molecules and excess body water from circulation [53]. After a prescribed dwell time, the fluid is removed and the process repeated four times within a 24-hour period for the average patient. The advantages of employing PD over hemodialysis (HD) include better quality of life for the patient with fewer hospitalizations and limitations on travel and work life. The frequency of PD versus HD worldwide varies considerably from 81% of all dialysis patients in Hong Kong to <10% in the United States and Germany and is related to reimbursement policies.

PD fluid employs a lactate-buffered glucose solution (1.36%, 2.27%, or 3.86%) as the most common type of osmotic agent, though icodextrin, a glucose polymer, is also available. Laboratory measures of urea (or urea nitrogen), creatinine, glucose or electrolytes in serum, urine, and the PD fluid are commonly requested in the management of PD patients, and the results entered into a software program designed to perform necessary calculations and interpretation of those results [54]. The peritoneal equilibration test (PET) is usually performed 4–6 weeks after initiation of PD with the goal to estimate the rate of small solute transport across the peritoneal membrane and the ultrafiltration capacity, which serves to characterize the patient's membrane transporter characteristics and guide the selection of an appropriate dialysate fluid glucose concentration and dwell time. Faster transport rates are observed when the PD creatinine to serum creatinine ratio after a 4-hour dwell time is largest. Smaller ultrafiltration volumes and slower transport rates are observed when the ratio of PD glucose after 4-hour dwell to PD glucose at baseline is greatest. PD adequacy is determined by collecting samples from a 24-hour dialysate and urine collection to measure urea and calculate Kt/V, where K is clearance of urea, t is dialysis time, and V is the volume of distribution of urea which is approximately equal to the patient's total body water. The result is multiplied by 7 to determine the weekly urea clearance with the goal to maintain weekly Kt/V >1.7. It should be measured at the same time as the PET is conducted and then quarterly or as clinically indicated.

The leading cause of complications for PD patients is peritonitis as a result of infection. The inflamed peritoneum becomes leaky whereby the osmotic gradient dissipates more rapidly and proteins cross more readily necessitating more frequent exchanges to maintain an adequate Kt/V. Antibiotics and/or heparin may be added to the dialysis fluid to treat infection and prevent the formation of fibrin.

Pericardial fluid

The pericardial cavity lies within the thoracic space, surrounding the heart and major vessels. Under normal conditions, 25–35 mL of fluid is present within that space to decrease friction and protect the organ from physical insults [55]. Pericardial effusions develop under pathological conditions when the rate of fluid formation exceeds the rate of removal. Pericardial effusions can be highly variable in presentation and may accompany a variety of pericardial disorders [55]. The onset can be acute (<1 week) to chronic (>3 months), vary in size and distribution, and remain essentially asymptomatic or cause significant hemodynamic instability [56]. The classic symptoms of effusion include shortness of breath, dyspnea on exertion, fever, chills, chest pain, and swelling. Cardiac tamponade is characterized by hypotension and increased pressure within the jugular vein. The physical exam often reveals neck vein distension, diminished heart sounds upon auscultation, and pulsus paradoxus—a phenomenon describing the reduction in radial pulse upon inspiration [56].

Pericardiocentesis is the procedure whereby a few milliliters to several liters of pericardial fluid is removed via percutaneous needle using echocardiographic guidance for diagnostic and/or therapeutic purposes [57]. Laceration or perforation of the myocardium or coronary vessels constitutes the major procedure-related complications, having a reported incidence rate of 1%–1.6% [56]. Pericardial fluid analysis is indicated according to the European Society of Cardiology guidelines in patients with cardiac tamponade, suspicion of infection or neoplasm, and in patients with moderate to large effusions of unknown etiology that have not responded to therapy [58]. Thus the analysis most often centers on the use of Gram stain and culture as well as the identification of malignant cells by cytological evaluations.

The differential diagnosis for pericardial effusion includes infection, connective tissue disease, neoplasm, metabolic causes, and idiopathic etiologies usually assumed to be viral infection [56]. Table 27.2 lists the most commonly reported causes of moderate and large pericardial effusions in several population-based studies. It is notable that in South Africa (and likely other parts of the world where tuberculosis infection is prevalent), infectious causes of pericardial effusion predominate, whereas Western nations report higher prevalence of idiopathic pericarditis, neoplastic, and iatrogenic causes [56]. In 60% of cases in which pericardial effusion is discovered, a known underlying disease exists and often can be treated without a diagnostic

TABLE 27.2 Prevalence and etiology of reported moderate and large pericardial effusions.

Study	Study period	*n* patients	Top three reported causes
United States	1993–2000	441	Iatrogenic 42%[a]
			Postoperative 28%
			Neoplastic 25%
Spain	1990–96	322	Idiopathic 29%
			Iatrogenic 16%
			Neoplastic 13%
France	1998–2002	204	Idiopathic 48%
			Infection 16%
			Neoplastic 15%
South Africa	1995–2001	233	Infection 72%
			Idiopathic 14%
			Neoplastic 9%
China	2007–09	140	Neoplastic 39%
			Infection 29%
			Idiopathic 9%
			Iatrogenic 9%

Notes: Refer to [56] and [60].
[a]*Postoperative causes and cardiac perforation from invasive procedure.*

pericardial fluid workup [56]. Inaccurate classification of pericardial effusions as transudative and exudative when applying classic Light's criteria (see Pleural fluid) has been reported. Therefore there is limited evidence that the measurement of most biochemical and tumor marker analytes in pericardial fluid adds any value to clinical management [61–65].

The most common infectious cause of pericardial effusions are viruses in the United States and Western European populations, whereas tuberculosis is a more frequent cause in developing countries [56]. Viral pericarditis is diagnosed based on clinical symptoms of pericarditis and elevated inflammatory markers, and it is often treated empirically with nonsteroidal antiinflammatory drugs (NSAIDs) or other antiinflammatory therapies. AFB staining and mycobacterium cultures are the mainstay for diagnosis of TB pericarditis; however, testing is limited by sensitivity and turnaround time of result availability. Biomarkers of infection have also been evaluated, and measurement of ADA in pericardial fluids has a reported 88% sensitivity, 83% specificity, and area under the receiver operator curve of 0.9539 using a cutoff >40 U/L by the Giusti method to identify tuberculous pericarditis in populations where the index of tuberculosis infection is sufficiently high [66]. For this reason, ADA measurement is best reserved for regions with moderate to high incidence of TB (approximately 50 cases per 100,000 or more) where the positive predictive value is increased with a lower rate of false positive results [67].

Malignant pericardial effusions are associated most commonly with lung carcinoma [68]. The fluid appears bloody more frequently compared to the serous appearance associated with other etiologies. Cytologic examination for malignant cells is the gold standard for diagnosis. The investigation should be performed on a sufficient volume of sample, at least 60 mL, to adequately rule out the presence of malignant or atypical cells suggestive of malignancy [69]. Dragoescu et al. reported a cohort of 128 specimens analyzed over 6 years in which cytology was 71% sensitive and 100% specific for detecting malignancy compared to pericardial biopsy which was 64% sensitive and 85% specific, although 87% of cases had a pathology-proven history of cancer at presentation [68]. In a separate cohort of patients presenting with large symptomatic malignant effusions with unrecognized cancer, cytology was only 51% sensitive, posing a diagnostic dilemma [70]. In cases of equivocal or suspected false negative cytology results, tumor marker analysis may be considered although such testing has reported limited success [4].

Synovial fluid

Synovial fluid is present in joint cavities and serves a number of important roles in maintaining joint health and mobility. It forms as a plasma ultrafiltrate, with additional components produced from synoviocytes. The synovial membrane lines the inner surface of the joint cavity. Type A synoviocytes are derived from monocytes and serve a phagocytic role in clearing cellular debris from the joint. Type B synoviocytes are fibroblasts that secrete substances, which help to lubricate and maintain the joint as discussed below. Small molecules (such as electrolytes) may filter freely into the synovial fluid, making their concentrations similar to those found in plasma. Size exclusion plays an important role in limiting larger molecules from filtering into the synovial fluid. For example, larger-molecular-weight proteins such as clotting factors typically do not enter the synovial fluid, whereas smaller proteins such as albumin may enter more easily.

Two important substances produced by synoviocytes are hyaluronic acid (HA) and lubricin. HA (or hyaluronan) is a high-molecular-weight polymer of repeating disaccharides (D-glucuronic acid and *N*-acetyl-D-glucosamine). HA increases synovial fluid viscosity, protects the surface of cartilage during stress and impact, and helps to maintain the joint space by influencing water and solute balance. Lubricin (also known as proteoglycan 4; PRG4) promotes joint lubrication and helps to protect articular cartilage. Together, these substances reduce friction and maintain joint health and mobility.

Normal and noninflammatory synovial fluid is typically clear or slightly yellow and somewhat viscous. Synovial fluid that is bloody or brownish in color may be consistent with trauma or recent and/or prior bleeding into the joint space (e.g., hemarthrosis or traumatic tap). In patients with unexplained bloody effusions, the possibility of coagulopathy (e.g., hemophilia) should be included in the differential diagnosis, particularly in pediatric patients. Cloudy, yellowish effusions may be consistent with inflammation and/or infection. Purulence is suggestive of infection, although cloudy-white effusions may also be observed in inflammatory or crystal-induced arthritis. While synovial fluid appearance is often suggestive of a particular etiology, it is not specific for individual diagnoses. Noninflammatory arthritis can be seen with conditions such as osteoarthritis (where protective cartilage is gradually lost from the ends of bone) or trauma. Inflammatory arthritis can have a variety of causes including rheumatoid arthritis, crystal arthritis, systemic lupus erythematosus, and other autoimmune disorders.

Symptoms of joint problems include pain, swelling, stiffness, or decreased range of motion. Septic joints may be accompanied by systemic symptoms such as fever. Joint problems can be localized to individual joints (e.g., knee osteoarthritis), present across many joints (e.g., rheumatoid arthritis), or may spread progressively to nearby joints (e.g., infectious arthritis). The acute versus chronic

nature of symptoms, as well the patient's medical history, often provide clues to underlying etiology.

The technique of synovial fluid collection is called arthrocentesis. Collection should be aseptic to prevent iatrogenic infection of otherwise noninfected joint spaces. Proper approach of the collection needle varies by the location of the joint to be aspirated [71]. Routine analysis of synovial fluid should include Gram stain, culture, crystal analysis, and cell count with WBC differential (to determine the percent of PMN leukocytes; %PMNs), as these tests are effective in helping to support the diagnosis of most inflammatory and noninflammatory diagnoses. Chemical analysis may provide additional information in certain cases. Manual tests (such as the string test for synovial fluid viscosity and the mucin test for clot detection) have limited clinical use.

WBC counts in normal synovial fluid specimens are usually in the range of 0−200 cells/μL. Increased WBC counts are typically seen in noninflammatory (0 to 2,000 cells/μL), inflammatory (2,000 to 100,000 cells/μL), and septic (10,000 to >100,000 cells/μL) joint disorders, although these are only approximate (and obviously overlapping) ranges, which serve as preliminary clues to an underlying category of disorder. Exceptions to these ranges are also quite common. For example, crystal synovitis may be observed with high WBC counts, whereas early septic arthritis may be observed with unexpectedly low WBC counts. Inflammatory causes of arthritis are typically associated with elevated %PMNs (>50%), while septic arthritis may have a %PMN of greater than 75%−90% [72]. Given these overlapping categorizations for WBC and %PMN, the importance of Gram stain and culture in helping to identify infectious etiologies is extremely important.

Crystal analysis should ideally be conducted using polarized microscopy. Polarized filters with a red compensator are used to detect birefringence, the double refraction of light when polarized rays hit certain crystalline materials. Many crystals (such as table salt, NaCl) are not birefringent. Monosodium urate (MSU) crystals, which are typically seen in gout, are needle-shaped in appearance and negatively birefringent. Calcium pyrophosphate dihydrate (CPPD) crystals, typically seen in pseudogout, have a rhomboid shape and are positively birefringent. When the long axis of a crystal is aligned in parallel to the slow wave of the microscope's red compensator, MSU crystals appear yellow while CPPD crystals appear blue. Other crystalline forms that may sometimes be observed in joint fluid include cholesterol, lipids, basic calcium phosphates/hydroxyapatite, calcium oxalate, and even corticosteroids after prior therapeutic injections. Acute attacks of gout and pseudogout are often treated with antiinflammatory medications (such as NSAIDs), colchicine, or corticosteroids. Prophylactic management of gout may include diet modification or medications such as xanthine oxidase inhibitors (e.g., allopurinol), uricosuric agents that increase uric acid excretion into the urine (e.g., probenicid), or enzymes that metabolize uric acid (e.g., uricase).

Early recognition and diagnosis of acute synovitis is essential to initiation of appropriate therapy, to minimize joint damage, and to maintain healthy joint function. Identification of septic arthritis is a primary concern. Common pathogens vary by patient age and underlying medical condition but include *Staphylococcus aureus*, *Streptococcus agalactiae* and *S. pyogenes*, *Neisseria gonorrhoeae*, *H. influenzae*, and more rarely Gram-negative bacteria such as *Pseudomonas aeruginosa*. Clues to disseminated gonococcal infection may include accompanying skin lesions and/or petechia, arthralgia, and arthritis. Risk for septic arthritis are increased in the presence of bacteremia, localized bacterial infections, joint implants, IV drug abuse, or chronic diseases associated with acute or chronic arthritis. Fastidious and atypical organisms may be difficult or impossible to identify by Gram stain or culture. Diagnosis of Lyme arthritis due to *Borrelia burgdorferi* is primarily serological (or molecular) but not conducted on joint fluid.

Chemical analysis of synovial fluids is sometimes ordered as supportive evidence for inflammatory and infectious (versus noninflammatory) causes of arthritis, but in general it is less clinically useful than routine cell count with differential, Gram stain, and culture. In normal synovial fluid, glucose concentrations are similar to those observed in fasting serum. Low synovial fluid glucose has been associated with septic arthritis or inflammation. In normal synovial fluid, total protein concentrations are typically 1/3 of a corresponding plasma specimen. Increased total protein can be observed with inflammation or damage to synovial membranes. Both glucose and protein have inadequate sensitivity and specificity to justify their routine clinical use.

Elevations in synovial fluid LDH have also been associated with both inflammatory and septic arthritis but again provide little benefit over the routine recommended tests. Increased lactate concentrations may also be observed with septic arthritis but are also not considered routine. Serum testing for inflammatory markers, such as rheumatoid factor, erythrocyte sedimentation rate, and C-reactive protein play a role in evaluating systemic inflammatory disease, although they are not considered clinically useful for most synovial fluids. Uric acid measurements may be useful in circumstances where microscopy is not available (or MSU crystals are not observed) but gout arthritis is strongly suspected [73].

Amniotic fluid

Amniotic fluid surrounds and protects a growing fetus within the uterus contained by a membrane known as the

amnion [69]. The volume of fluid contained within the amnion increases from approximately 50 mL present at 12 weeks gestation to approximately 1 L at term. Polyhydramnios is a condition where the volume of amniotic fluid is abnormally increased, while oligohydramnios is a condition of abnormally decreased amniotic fluid volume. The composition of amniotic fluid varies throughout gestation. Early in gestation the amniotic fluid is produced by the amnion and placenta, representing a plasma ultrafiltrate. As gestation progresses, the fetal lungs and kidneys develop and fluid exchange within the amnion gives rise to increasing concentrations of metabolic byproducts including creatinine, urea, and uric acid. These waste products are removed by exchange with maternal serum. Fetal respiratory action develops during the latter weeks of gestation (≥32 weeks) causing fetal pulmonary surfactant molecules to be released into the amniotic fluid.

Amniocentesis is a percutaneous procedure whereby a needle is inserted transabdominally, under continuous US guidance, into an unoccupied area of the amniotic cavity to remove 10−20 mL of amniotic fluid using a syringe [21]. The first drops of fluid are typically wasted as they may contain maternal cells and could be mistaken as mosaicism in cytogenetic studies [74]. The volume of fluid that can be safely removed is approximately 1 mL for each gestational week (e.g., 17 mL at 17 weeks). Diagnostic amniocentesis performed during 14−18 weeks of gestation may be conducted to perform genetic studies if there is suspicion for increased risk of chromosomal abnormalities based on medical history, US examination, or after first or second trimester maternal serum screening. In limited cases where maternal serum screening tests reveal increased risk of open neural tube defects and high-resolution US findings are inconclusive, measurement of AFP and acetylcholinesterase activity in amniotic fluid is recommended. Refer to Chapter 42, Laboratory testing in pregnancy, for a complete description of lab testing for fetal anomalies.

Amniocentesis is performed during 20−42 weeks of gestation to assess fetal distress. Intraamniotic infection or chorioamnionitis is an acute inflammation of the fetal membranes commonly caused by bacterial infection prompting an inflammatory response leading to labor and term or preterm birth [75]. Chorioamnionitis may be symptomatic (clinical) or asymptomatic (histological), occurring most often during prolonged labor or as a consequence of membrane rupture as bacteria have greater opportunity to ascend the lower genital tract to colonize the uterus. Prompt diagnosis and treatment for clinical chorioamnionitis is critical to avoid maternal and fetal morbidity and mortality. The symptoms are not universally agreed upon but include maternal fever and maternal or fetal tachycardia, uterine fundal tenderness, or purulent amniotic fluid. Histopathologic chorioamnionitis occurs when there is microscopic evidence of infection or inflammation of the placenta or chorion which may be accompanied by symptoms of infection or remain asymptomatic and only diagnosed after delivery. Hemolytic disease of the newborn also causes fetal distress due to alloimmunization and Rh incompatibility between mother and fetus. Amniotic fluid may be collected to undergo PCR analysis to genotype an at-risk fetus as well as measure bilirubin to assess the extent of hemolysis and need for intrauterine transfusion.

Historically after week 32, fetal lung maturity (FLM) was evaluated to make decisions to delay an elective delivery while corticosteroids were administered to promote lung development if necessary. Identification of pulmonary surfactants (lamellar bodies, lecithin, sphingomyelin, and phosphatidyl glycerol) in amniotic fluid is supportive evidence that the fetal lungs have sufficiently matured and the risk of developing respiratory distress syndrome is low. However, more recent clinical guidelines favor using gestational age to assess fetal maturity as the assessment of FLM did not improve outcomes and nonmedically indicated late-preterm and early-term deliveries are strongly discouraged due to the increased risk of fetal morbidity and mortality [59]. Laboratories have noted declining volumes of FLM testing in recent years as a likely signal of this trend [76]. Refer to Chapter 42, Laboratory testing in pregnancy, for an overview of FLM.

Therapeutic amniocentesis may be performed in cases of polyhydramnios. The complications associated with amniocentesis include rupture of membranes, fetal injury, infection, and fetal loss. Fetal loss related to amniocentesis typically occurs in the 4 weeks following the procedure at a rate of 1 in 300 to 1 in 500 [77]. Temporary leakage of fluid occurs more frequently (1.7%); however, it is generally associated with normal pregnancy outcomes.

Feces

Feces (also known as stool) is the waste product of digestion. Fecal analysis is typically conducted in the evaluation of diarrheal disease (frequent passage of loose, watery stool), suspected gastrointestinal infection with pathogens (e.g., bacteria, viruses, or parasites), evaluation for malabsorptive or maldigestive disorders, and for the detection of occult blood loss that may be associated with gastrointestinal bleeding or malignancy. Stool examination, as well as testing for fecal occult blood, malignancy, and malabsorptive disorders, is covered in Chapter 32, Gastrointestinal disorders.

Collection of fecal specimens is an unpleasant task for most patients and is often associated with poor compliance. Proper collection includes avoiding contamination with urine and toilet water. Handwashing after collection

is important to prevent pathogen spread. In some cases, multiple stool collections over several days may be required. Preservatives may be recommended or required for some fecal tests (e.g., detection of ova and parasites; O&P) but are incompatible with other analyses such as bacterial culture. Many laboratories have initiated specimen acceptability requirements for some fecal testing. For example, a well-formed stool specimen from an inpatient (hospitalized for several days) may be rejected for routine O&P or bacterial culture, as neither the specimen nor the medical scenario is consistent with ongoing diarrheal disease.

Routine fecal bacterial culture protocols are often designed to detect *Salmonella*, *Shigella*, and *Campylobacter* species. Many laboratories also include processes to detect *E. coli* O157:H7 and Shiga-toxin producing strains of *E. coli*. Bacterial identification may also include *Yersinia*, *Vibrio*, *Aeromonas*, *Plesiomonas*, and *Staphylococcus* species.

Both antigen and NAT have proven to be extremely valuable in fecal analyses. Production of *Clostridium difficile* toxin may be detected by enzyme immunoassay (EIA), PCR, or cell cytotoxicity assays. *E. coli* Shiga-like toxins may also be detected by EIA. *Cryptosporidium* EIA may provide supportive information along with O&P. Stool viral cultures may detect a number of pathogenic viruses, although specific viral antigen testing (e.g., for rotavirus or adenovirus serotypes 40 and 41) provides more rapid turnaround times. Multiplex NAT platforms may reduce the time for identification of bacterial, viral, and parasitic species.

The cause of chronic diarrhea (symptoms lasting >4 weeks) may be investigated using fecal electrolyte and osmolality studies to classify diarrhea as secretory (i.e., epithelial transport-mediated) or osmotic (i.e., due to the presence of poorly absorbed solutes in the GI tract). A Fecal Osmotic Gap is typically calculated as:

$$\text{Fecal osmotic gap} = \text{Fecal osmolality} - 2 \times \left([\text{Na}]_{\text{fecal}} + [\text{K}]_{\text{fecal}}\right)$$

An average fecal osmolality of 290 mOsm/kg is often used in the calculation; alternatively, fecal osmolality may be measured with an osmometer. A low fecal osmotic gap (<50 mOsm/kg) with an elevated fecal sodium is consistent with secretory diarrhea or use of sodium-containing laxatives. A high osmotic gap (>100 mOsm/kg) with a low fecal sodium is typically consistent with osmotic diarrhea. Elevated fecal magnesium concentrations provide evidence for use of magnesium-containing laxatives.

A variety of additional tests have been used to identify inflammatory, digestive, malabsorptive, and even systemic disorders. The detection of fecal leukocytes by stool microscopy is suggestive of GI inflammation or infection. Fecal lactoferrin and fecal calprotectin (substances produced by neutrophils) have both been used to differentiate inflammatory bowel disease (IBD) from irritable bowel syndrome (IBS) or other noninflammatory gastrointestinal disorders. Fecal alpha-1-antitrypsin (A1A) clearance measurements, which compare A1A concentration in 24-hour stool collections to serum, have been used to identify protein-losing enteropathies. Fecal bile acid measurements are essential to the identification of bile acid malabsorption. Increased fecal reducing substances (sugars) are suggestive of disorders in carbohydrate absorption or metabolism. Finally, fecal porphyrin testing is critical to the differentiation of variegate porphyria and hereditary coproporphyria from acute intermittent porphyria.

References

[1] Clinical Laboratory Improvement Amendments of 1988. Laboratory requirements, 42 C.F.R. § 493.1253(b) (2011).

[2] D.R. Block, D.D.H. Franke, Quick Guide to Body Fluid Testing, AACC, Washington, DC, 2015.

[3] D.R. Block, L.J. Ouverson, C.A. Wittwer, A.K. Saenger, N.A. Baumann, An approach to analytical validation and testing of body fluid assays for the automated clinical laboratory, Clin. Biochem. 58 (2018) 44–52.

[4] D.R. Block, A. Algeciras-Schimnich, Body fluid analysis: clinical utility and applicability of published studies to guide interpretation of today's laboratory testing in serous fluids, Crit. Rev. Clin. Lab. Sci. 50 (4–5) (2013) 107–124.

[5] W.E. Owen, M.L. Thatcher, K.J. Crabtree, R.W. Greer, F.G. Strathmann, J.A. Straseski, et al., Body fluid matrix evaluation on a Roche cobas 8000 system, Clin. Biochem. 48 (13–14) (2015) 911–914.

[6] M.J. Lin, C. Hoke, R. Dlott, T.S. Lorey, D.N. Greene, Performance specifications of common chemistry analytes on the AU series of chemistry analyzers for miscellaneous body fluids, Clin. Chim. Acta. 426 (2013) 121–126.

[7] W.E. Owen, J.J.H. Hunsaker, J.R. Genzen, Alpha-fetoprotein in pericardial, peritoneal, and pleural fluids: a body fluid matrix evaluation, Clin. Biochem. 56 (2018) 109–112.

[8] S.Y. Lo, N.H. Saifee, B.O. Mason, D.N. Greene, Filling in the gaps with non-standard body fluids, Pract. Lab. Med. 5 (2016) 24–31.

[9] D.S. Jacobs, W.R. DeMott, D.K. Oxley, Jacobs & DeMott Laboratory Test Handbook., fifth ed., Lexi-Comp, Hudson, OH, 2001.

[10] N. Rifai, A.R. Horvath, C.T. Wittwer, Tietz Textbook of Clinical Chemistry and Molecular Diagnostics., sixth ed., Elsevier, St. Louis, Mo, 2018.

[11] J.W. Ironside, D.L. Ritchie, M.W. Head, Prion diseases, Handb. Clin. Neurol. 145 (2017) 393–403.

[12] D.W. Colby, S.B. Prusiner, Prions, Cold Spring Harb. Perspect. Biol. 3 (1) (2011) a006833.

[13] R. Atarashi, K. Satoh, K. Sano, T. Fuse, N. Yamaguchi, D. Ishibashi, et al., Ultrasensitive human prion detection in cerebrospinal fluid by real-time quaking-induced conversion, Nat. Med. 17 (2) (2011) 175–178.

[14] Pizella, N. The proper handling of CJD-infected patient samples in the pathology laboratory. 2018. <https://www.mlo-online.com/disease/infectious-disease/article/13009473/the-proper-handling-of-cjdinfected-patient-samples-in-the-pathology-laboratory>. (accessed 11.09.19).

[15] WHO Infection Control Guidelines for Transmissible Spongiform Encephalopathies. Report of a WHO Consultation. 1999.
[16] S.A. Sahn, Getting the most from pleural fluid analysis, Respirology. 17 (2) (2012) 270−277.
[17] J.M. Porcel, R.W. Light, Diagnostic approach to pleural effusion in adults, Am. Fam. Physician. 73 (7) (2006) 1211−1220.
[18] C.E. Gordon, D. Feller-Kopman, E.M. Balk, G.W. Smetana, Pneumothorax following thoracentesis: a systematic review and meta-analysis, Arch. Intern. Med. 170 (4) (2010) 332−339.
[19] J. Hussong, C. Kjeldsberg (Eds.), Kjeldsberg's Body Fluid Analysis, American Society for Clinical Pathology, Chicago, IL, 2015.
[20] J.M. Porcel, Identifying transudates misclassified by light's criteria, Curr. Opin. Pulm. Med. 19 (4) (2013) 362−367.
[21] N.A. Brunzel, Fundamentals of Urine & Body Fluid Analysis., third ed., Elsevier/Saunders, St. Louis, Mo, 2013.
[22] L. Mundt, K. Shanahan, Graff's Textbook of Urinalysis and Body Fluids., third ed., Wolters Kluwer, Philadelphia, PA, 2016.
[23] S.K. Strasinger, M.S. Di Lorenzo, Urinalysis and Body Fluids., sixth ed., F.A. Davis Company, Philadelphia, PA, 2014.
[24] R.W. Light, M.I. Macgregor, P.C. Luchsinger, W.C. Ball Jr., Pleural effusions: the diagnostic separation of transudates and exudates, Ann. Intern. Med. 77 (4) (1972) 507−513.
[25] M.C. Houston, Pleural fluid pH: diagnostic, therapeutic, and prognostic value, Am. J. Surg. 154 (3) (1987) 333−337.
[26] J.S. Hackbarth, K. Murata, W.M. Reilly, A. Algeciras-Schimnich, Performance of CEA and CA19-9 in identifying pleural effusions caused by specific malignancies, Clin. Biochem. 43 (13−14) (2010) 1051−1055.
[27] C. Hooper, Y.C. Lee, N. Maskell, Group BTSPG, Investigation of a unilateral pleural effusion in adults: British Thoracic Society Pleural Disease Guideline 2010, Thorax. 65 (Suppl 2) (2010) ii4−17.
[28] C.M. Moore, D.H. Van Thiel, Cirrhotic ascites review: pathophysiology, diagnosis and management, World J. Hepatol. 5 (5) (2013) 251−263.
[29] J. Scallan, V.H. Huxley, R.J. Korthuis, Capillary Fluid Exchange: Regulation, Functions, and Pathology, Morgan & Claypool Life Sciences, San Rafael, CA, 2010.
[30] L.R. Roberts, P.S. Kamath, Ascites and hepatorenal syndrome: pathophysiology and management, Mayo Clin. Proc. 71 (9) (1996) 874−881.
[31] C.M. Grabau, S.F. Crago, L.K. Hoff, J.A. Simon, C.A. Melton, B. J. Ott, et al., Performance standards for therapeutic abdominal paracentesis, Hepatology. 40 (2) (2004) 484−488.
[32] B.A. Runyon, AASLD. Introduction to the revised American Association for the study of liver diseases practice guideline management of adult patients with ascites due to cirrhosis 2012, Hepatology. 57 (4) (2013) 1651−1653.
[33] A.C. Tarn, R. Lapworth, Biochemical analysis of ascitic (peritoneal) fluid: what should we measure, Ann. Clin. Biochem. 47 (Pt 5) (2010) 397−407.
[34] B.A. Runyon, A.A. Montano, E.A. Akriviadis, M.R. Antillon, M. A. Irving, J.G. McHutchison, The serum-ascites albumin gradient is superior to the exudate-transudate concept in the differential diagnosis of ascites, Ann. Intern. Med. 117 (3) (1992) 215−220.
[35] W. Hou, A.J. Sanyal, Ascites: diagnosis and management, Med. Clin. North. Am. 93 (4) (2009) 801−817.
[36] B.A. Runyon, A.P.G. Committee, Management of adult patients with ascites due to cirrhosis: an update, Hepatology. 49 (6) (2009) 2087−2107.
[37] E.A. Akriviadis, B.A. Runyon, Utility of an algorithm in differentiating spontaneous from secondary bacterial peritonitis, Gastroenterology. 98 (1) (1990) 127−133.
[38] A.Q. Farias, O.M. Silvestre, G. Garcia-Tsao, L.F. da Costa Seguro, D.F. de Campos Mazo, F. Bacal, et al., Serum B-type natriuretic peptide in the initial workup of patients with new onset ascites: a diagnostic accuracy study, Hepatology. 59 (3) (2014) 1043−1051.
[39] F.A. Rochling, R.K. Zetterman, Management of ascites, Drugs. 69 (13) (2009) 1739−1760.
[40] E.J. Kaleta, N.V. Tolan, K.A. Ness, D. O'Kane, A. Algeciras-Schimnich, CEA, AFP and CA 19-9 analysis in peritoneal fluid to differentiate causes of ascites formation, Clin. Biochem. 46 (9) (2013) 814−818.
[41] M. Ahadi, S. Tehranian, B. Memar, H. Vossoughinia, M. Salari, E. Eskandari, et al., Diagnostic value of carcinoembryonic antigen in malignancy-related ascites: systematic review and meta-analysis, Acta Gastroenterol. Belg. 77 (4) (2014) 418−424.
[42] B. Lizaola, A. Bonder, H.D. Trivedi, E.B. Tapper, A. Cardenas, Review article: the diagnostic approach and current management of chylous ascites, Aliment. Pharmacol. Ther. 46 (9) (2017) 816−824.
[43] A. Cardenas, S. Chopra, Chylous ascites, Am. J. Gastroenterol. 97 (8) (2002) 1896−1900.
[44] M.A. Thaler, A. Bietenbeck, C. Schulz, P.B. Luppa, Establishment of triglyceride cut-off values to detect chylous ascites and pleural effusions, Clin. Biochem. 50 (3) (2017) 134−138.
[45] Y.C. Shen, T. Wang, L. Chen, T. Yang, C. Wan, Q.J. Hu, et al., Diagnostic accuracy of adenosine deaminase for tuberculous peritonitis: a meta-analysis, Arch. Med. Sci. 9 (4) (2013) 601−607.
[46] G. Giusti, B. Galanti, Colorimetric method, in: H.U. Bergmeyer (Ed.), Methods of Enzymatic Analysis., first ed, Verlag Chemie, Weinheim, Germany, 1984, pp. 315−323.
[47] P. Darwin, E. Goldberg, L. Uradomo, Jackson Pratt drain fluid-to-serum bilirubin concentration ratio for the diagnosis of bile leaks, Gastrointest. Endosc. 71 (1) (2010) 99−104.
[48] A.T. DeBenedet, J.M. Scheiman, G.H. Elta, B.J. Elmunzer, Peritoneal fluid bilirubin to serum bilirubin ratio for the diagnosis of bile leaks in orthotopic liver transplant recipients, Dig. Dis. Sci. 58 (10) (2013) 3044−3048.
[49] H. Engel, D.J. Bac, R. Brouwer, B.G. Blijenberg, J. Lindemans, Diagnostic analysis of total protein, albumin, white cell count and differential in ascitic fluid, Eur. J. Clin. Chem. Clin Biochem. 33 (4) (1995) 239−242.
[50] S. Thompson, L.E. Ding, Underestimation of the serum ascites albumin gradient by the bromocresol purple method of albumin measurement, Intern. Med. J. 48 (11) (2018) 1412−1413.
[51] T.M. Lehto, P. Leskinen, P. Hedberg, T.E. Vaskivuo, Evaluation of the Sysmex XT-4000i for the automated body fluid analysis, Int. J. Lab. Hematol. 36 (2) (2014) 114−123.
[52] O. Riggio, S. Angeloni, Ascitic fluid analysis for diagnosis and monitoring of spontaneous bacterial peritonitis, World J. Gastroenterol. 15 (31) (2009) 3845−3850.
[53] B. Rippe, Peritoneal dialysis, in: R.J. Johnson, J. Feehally, J. Floege (Eds.), Comprehensive Clinical Nephrology., fifth ed., Saunders/Elsevier, Philadelphia, PA, 2010.

[54] Hemodialysis Adequacy, Work Group. Clinical practice guidelines for hemodialysis adequacy, update 2006, Am. J. Kidney Dis. 48 (Suppl 1) (2006) S2–S90.
[55] S.J. Hutchison, Pericardial Diseases: Clinical Diagnostic Imaging Atlas, Saunders/Elsevier, Philadelphia, PA, 2009.
[56] M. Imazio, Y. Adler, Management of pericardial effusion, Eur. Heart J. 34 (16) (2013) 1186–1197.
[57] H. Mallemat, S. Tewelde, in: J. Roberts (Ed.), Roberts and Hedges' Clinical Procedures in Emergency Medicine, Elsevier Saunders, Philadelphia, PA, 2014.
[58] B. Maisch, P.M. Seferovic, A.D. Ristic, R. Erbel, R. Rienmuller, Y. Adler, et al., Guidelines on the diagnosis and management of pericardial diseases executive summary: The Task Force on the Diagnosis and Management of Pericardial Siseases of the European Society of Cardiology, Eur. Heart J. 25 (7) (2004) 587–610.
[59] American College of Obstetricians and Gynecologists, Committee Opinion No. 561. Nonmedically indicated early-term deliveries, Obstet. Gynecol. 121 (4) (2013) 911–915.
[60] T.S. Tsang, M. Enriquez-Sarano, W.K. Freeman, M.E. Barnes, L. J. Sinak, B.J. Gersh, et al., Consecutive 1127 therapeutic echocardiographically guided pericardiocenteses: clinical profile, practice patterns, and outcomes spanning 21 years, Mayo Clin. Proc. 77 (5) (2002) 429–436.
[61] S. Ben-Horin, I. Bank, A. Shinfeld, E. Kachel, V. Guetta, A. Livneh, Diagnostic value of the biochemical composition of pericardial effusions in patients undergoing pericardiocentesis, Am. J. Cardiol. 99 (9) (2007) 1294–1297.
[62] S. Ben-Horin, A. Shinfeld, E. Kachel, A. Chetrit, A. Livneh, The composition of normal pericardial fluid and its implications for diagnosing pericardial effusions, Am. J. Med. 118 (6) (2005) 636–640.
[63] L.J. Burgess, Biochemical analysis of pleural, peritoneal and pericardial effusions, Clin. Chim. Acta. 343 (1–2) (2004) 61–84.
[64] L.J. Burgess, H. Reuter, J.J. Taljaard, A.F. Doubell, Role of biochemical tests in the diagnosis of large pericardial effusions, Chest. 121 (2) (2002) 495–499.
[65] S. Akyuz, E. Arugaslan, A. Zengin, T. Onuk, U.S. Ceylan, B. Yaylak, et al., Differentiation between transudate and exudate in pericardial effusion has almost no diagnostic value in contemporary medicine, Clin. Lab. 61 (8) (2015) 957–963.
[66] F.F. Tuon, M.N. Litvoc, M.I. Lopes, Adenosine deaminase and tuberculous pericarditis--a systematic review with meta-analysis, Acta Trop. 99 (1) (2006) 67–74.
[67] F.F. Tuon, V.I. Silva, G.M. Almeida, L. Antonangelo, Y.L. Ho, The usefulness of adenosine deaminase in the diagnosis of tuberculous pericarditis, Rev. Inst. Med. Trop. Sao Paulo. 49 (3) (2007) 165–170.
[68] E.A. Dragoescu, L. Liu, Pericardial fluid cytology: an analysis of 128 specimens over a 6-year period, Cancer Cytopathol. 121 (5) (2013) 242–251.
[69] L.M. Rooper, S.Z. Ali, M.T. Olson, A minimum volume of more than 60mL is necessary for adequate cytologic diagnosis of malignant pericardial effusions, Am. J. Clin. Pathol. 145 (1) (2016) 101–106.
[70] S. Ben-Horin, I. Bank, V. Guetta, A. Livneh, Large symptomatic pericardial effusion as the presentation of unrecognized cancer: a study in 173 consecutive patients undergoing pericardiocentesis, Med. (Baltim). 85 (1) (2006) 49–53.
[71] P. Courtney, M. Doherty, Joint aspiration and injection and synovial fluid analysis, Best. Pract. Res. Clin. Rheumatol. 27 (2) (2013) 137–169.
[72] Sholter D., Russell A. Synovial Fluid Analysis. UpToDate. Shmerling RH and Calderwood SB, Section Editors. Curtis MR, Deputy Editor. <http://www.uptodate.com>. (accessed 9.15.18).
[73] B. Vaidya, M. Bhochhibhoya, S. Nakarmi, Synovial fluid uric acid level aids diagnosis of gout, Biomed. Rep. 9 (1) (2018) 60–64.
[74] Ghidini A. Diagnostic Amniocentesis. UpToDate. Wilkins-Haug L and Levine D, Section Editors. Barss VA, Deputy Editor. <http://www.uptodate.com>. (accessed 9.15.18).
[75] A.T. Tita, W.W. Andrews, Diagnosis and management of clinical chorioamnionitis, Clin. Perinatol. 37 (2) (2010) 339–354.
[76] M.L. Yarbrough, D.G. Grenache, A.M. Gronowski, Fetal lung maturity testing: the end of an era, Biomark Med. 8 (4) (2014) 509–515.
[77] American College of Obstetricians and Gynecologists, Practice Bulletin No. 88. Invasive prenatal testing for aneuploidy, Obstet. Gynecol. 110 (6) (2007) 1459–1467.

Self-assessment questions

1. Exudates are fluid accumulations that are most often due to membrane damage, infection, inflammation, or malignancy.
   a. true
   b. false
2. Which of the following laboratory findings provides supportive evidence that a pleural fluid is an exudate:
   a. fluid LDH activity that is $<2/3$ the upper limit of a normal serum LDH activity
   b. fluid-to-serum LDH ratio of <0.6
   c. fluid-to-serum protein ratio of >0.5
   d. serum-to-fluid albumin gradient (serum albumin minus fluid albumin) >1.2 g/dL
   e. serum-to-fluid protein gradient (serum protein minus fluid protein) >3.1 g/dL
3. A SAAG $\geq$ 1.1 g/dL and a total protein >2.5 g/dL is most consistent with which of the following etiologies of the ascites:
   a. cirrhosis
   b. heart failure
   c. malignancy
   d. pancreatitis
   e. tuberculosis
4. The procedure to collect pleural fluid is known as:
   a. amniocentesis
   b. arthrocentesis
   c. paracentesis
   d. pericardiocentesis
   e. thoracentesis
5. The presence of CSF in a nasal fluid specimen can be evaluated by which of the following tests:
   a. alpha-1 antitrypsin
   b. alpha-2 macroglobulin
   c. beta-2 glycoprotein
   d. beta-2 transferrin
   e. beta-2 microglobulin
6. Which of the following fecal tests would be most useful in differentiating IBD from IBS:
   a. fecal calprotectin
   b. fecal fat
   c. fecal *Helicobacter pylori* antigen
   d. fecal porphyrins
   e. fecal reducing substances

Answers

1. a
2. c
3. b
4. e
5. d
6. a

Chapter 28

Lipids and lipoproteins

Leslie J. Donato and Jeffrey W. Meeusen
Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Discuss the mechanisms of lipid and lipoprotein production and metabolism.
- Explain the clinical testing methods used in the measurement of lipids and lipoproteins.
- Describe dyslipoproteinemias in terms of laboratory aberrations and clinical manifestations.

Background

In a clinical context, the term "lipids" almost invariably means cholesterol. This focus on a single molecule is due to the extensively studied association between cholesterol and cardiovascular disease (CVD). Despite this attention, cholesterol is not the most prevalent lipid carried in blood plasma nor is it directly pathological. In fact, considerable effort has been made to promulgate the concepts of "good cholesterol" versus "bad cholesterol," which will be discussed further below.

Clinical chemists know that lipids are a class of molecules with poor aqueous solubility. A small number of lipids bind directly with albumin or other globulins; however, most blood lipids are incorporated into lipoproteins. A lipoprotein is a heterogeneous particle composed of proteins and lipids. Solubility is maintained by locating apolipoproteins and polar lipids [e.g., phospholipids, sphingolipids, fatty acids (FAs), and cholesterol] near the outer surface of the particle. Conversely, nonpolar lipids such as triglycerides and cholesteryl esters are carried in the lipoprotein core. The primary role of lipoproteins is to redistribute lipids between organs and tissues via the bloodstream.

Assessment of plasma lipoproteins can be informative in the clinical evaluation of a patient. Abnormal lipoprotein distributions are often indicative of underlying disease, such as liver cholestasis, inborn errors of metabolism, or nutritional imbalances. Furthermore, abnormal plasma concentrations of lipoproteins may have severe consequences, including pancreatitis, neurological deficits, and CVD.

Clinically relevant lipids

Cholesterol is a tetracyclic alcohol that can be produced de novo by all animal cells (Fig. 28.1). Cellular membranes incorporate cholesterol as a structural component, where it provides increased strength without rigidity. Cholesterol is also a precursor for many essential compounds, including vitamin D, bile salts, and hormones.

FAs are carboxylic acids with a variable length hydrocarbon chain. FAs are synthesized via glycolysis and used as a means of energy storage with increased density as compared with glycogen or glucose. The hydrocarbon chain of FAs contains an even number of carbons (typically 4−26) that can be saturated (without double bonds) or unsaturated (with double bonds). FAs are often components of more complex lipids such as triglycerides, cholesteryl esters, and sphingolipids.

Cholesteryl esters are extremely hydrophobic and account for 60%−80% of all plasma cholesterol. Esterification of three FAs with a single glycerol molecule results in the production of a triglyceride molecule, which is the primary energy storage source in animals. Phospholipids are the product of esterification by two FAs and an inorganic phosphate with glycerol. Serine FA esters are precursors for sphingolipids (Fig. 28.1). Nonesterified FAs are also referred to as "free" fatty acids (FFA), and are generated when released from triglycerides and other esters following enzymatic lipolysis.

Lipoprotein metabolism

Apolipoprotein B lipoproteins and triglyceride redistribution

The role of apolipoprotein B (apoB)-lipoproteins is to redistribute triglycerides between tissues. ApoB-lipoproteins begin as triglyceride-enriched lipoproteins (TRLs) synthesized in the liver or intestines. TRLs are converted into smaller lipoproteins in circulating plasma through the enzymatic lipolysis of triglycerides into FFAs (Fig. 28.2).

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00028-4

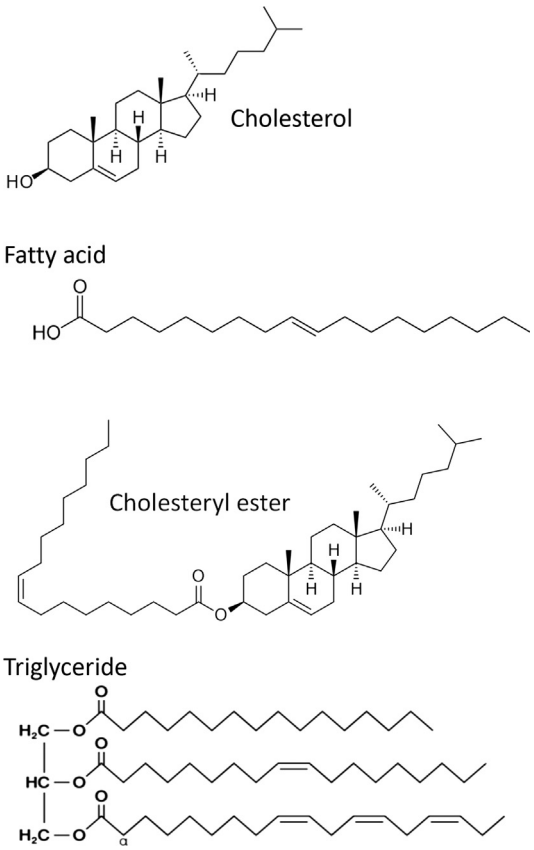


FIGURE 28.1 Common biological lipids.

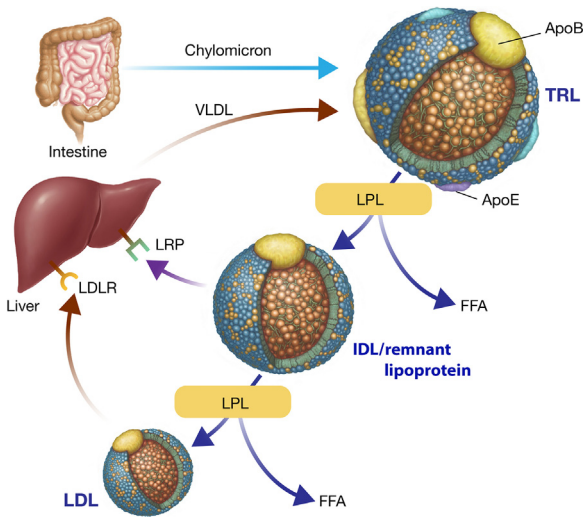


FIGURE 28.2 Apolipoprotein B metabolism. Triglyceride rich lipoproteins are generated by the intestines (chylomicrons: 75−1200-nm diameter) or liver (very-low-density lipoproteins: 30−80-nm diameter). Lipoprotein lipase activity in blood capillaries converts triglycerides in the lipoprotein core into free-fatty acids. The smaller and denser intermediate-density lipoproteins (from very-low-density lipoproteins) or remnant lipoproteins (from chylomicrons) are taken up by the liver via low-density lipoprotein receptors or low-density lipoprotein receptor-related protein. Low-density lipoproteins are apolipoprotein B-lipoproteins produced from very-low-density lipoproteins/intermediate-density lipoproteins and are too small to carry apolipoprotein E so are cleared only via low-density lipoprotein receptor.

ApoB is required for the synthesis of very-low-density lipoproteins (VLDLs)[1] by the liver to start the "intrinsic" lipid metabolism pathway employed for lipids in systemic circulation. A truncated apoB isoform (apoB-48) is synthesized in the intestines during the formation of chylomicrons to initiate the "extrinsic" lipid metabolism pathway from dietary lipids. Both intestinal apoB and hepatic apoB are encoded by a single gene and share a common N-terminal sequence. The shorter apoB-48 protein is produced by RNA editing of the full apoB transcript, resulting in the creation of a stop codon, and early translation termination. In VLDL and chylomicrons, as well as remnant lipoproteins (RLPs), apoB is the primary structural component around which the lipid particle is formed (Fig. 28.2).

In addition to apoB, VLDL and chylomicrons contain apolipoprotein E (apoE), apolipoprotein C-II (apoC-II), and apolipoprotein C-III (apoC-III). ApoE, apoC-II, and apoC-III are exchangeable apolipoproteins, meaning they can be added and removed from the lipoprotein. In this way, any given lipoprotein may contain several copies of apoE, apoC-II, and apoC-III, or none at all. ApoB is unique in that it is a nonexchangeable structural component of the lipoprotein and all apoB-lipoproteins always contain exactly one apoB (Fig. 28.2).

Lipoprotein lipase (LPL) expressed on the capillary endothelium in peripheral tissues converts triglycerides into FFAs. Tissues with a high demand for lipids, such as skeletal muscle and adipose, have increased capillary expression of LPL. The rate of lipolysis is regulated by apoC-II and apoC-III. ApoC-II is an essential cofactor for LPL, while apoC-III is a potent inhibitor of LPL. As LPL converts triglycerides into FFAs, which transfer into the tissues, the lipoprotein becomes smaller and increases in density. VLDL RLPs are referred to as intermediate-density lipoproteins (IDL), while chylomicron remnants are referred to as RLPs.

Hepatic lipases expressed on hepatocytes further reduce triglyceride content and aid in the conversion of TRLs into IDL and RLP. Under normal conditions, IDL and RLP are removed from circulation via two hepatic receptor proteins: the low-density lipoprotein receptor (LDLR) and LDLR-related protein (LRP). Both LDLR and LRP have a high affinity for apoE. Newly synthesized lipoproteins are large, with a diameter of 30−80 nm for VLDL and 75−1200 nm for chylomicrons. This is

1. Lipoprotein nomenclature is based on early studies that fractionated serum lipids by density.

noteworthy because the receptor domain of apoE is masked in large lipoproteins and only becomes available for binding once lipoproteins are reduced in size. This prevents immediate reuptake of newly synthesized lipoproteins by hepatocytes.

ApoB also has some affinity for LDLR, but it is much weaker than apoE. Under normal physiologic conditions, apoE-facilitated clearance is highly efficient and prevents measureable accumulation of IDL or chylomicron remnants. Unfortunately, this clearance may be incomplete. Some lipoproteins lose nearly all triglycerides and all apoE prior to removal from circulation. The result is a low-density lipoprotein (LDL). LDL is rich in cholesterol and contains no apolipoproteins other than apoB. Clearance of LDL relies on the low-affinity interaction between the LDLR and a single apoB ligand per particle (Fig. 28.2); consequently, the in vivo half-life of LDL is significantly longer than other apoB-lipoproteins.

Apolipoprotein A-I lipoproteins and reverse cholesterol transport

Apolipoprotein A-I (apoA1) is the structural protein of high-density lipoproteins (HDLs). In lipid metabolism, the role of apoA1 (and HDL) is to facilitate the removal of excess cholesterol from peripheral tissues. This process is referred to as reverse cholesterol transport (Fig. 28.3). ApoA1 is synthesized by both hepatocytes and enterocytes, as is apoB. However, while apoB-lipoproteins are fully assembled prior to secretion, apoA1 is released directly into circulation. This lipid-poor apoA1 is referred to as nascent HDL (also called preß-HDL, due to its electrophoretic migration pattern).

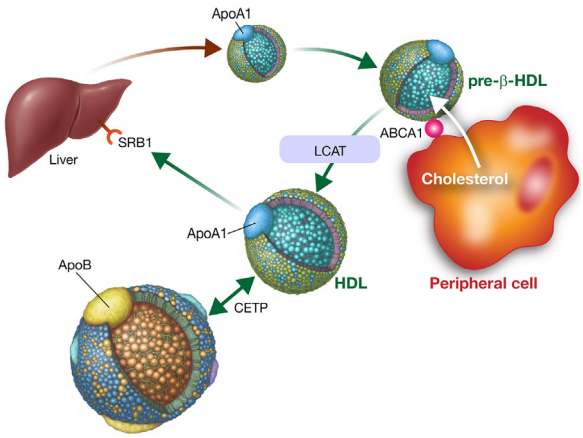


FIGURE 28.3 Reverse cholesterol transport pathway. Lipid-poor Apolipoprotein A-I is secreted by the liver and grows into a pre-ß-high-density lipoprotein as cholesterol from peripheral cells is transferred via adenosine triphosphate binding cassette A1. The pre-ß-high-density lipoprotein is converted into a mature high-density lipoprotein by the activity of lecithin:cholesterol acyltransferase. Cholesterol and triglycerides are exchanged between the mature high-density lipoproteins and the apolipoprotein B-lipoproteins via cholesterol ester transfer protein. High-density lipoprotein is cleared by the liver via scavenger receptor class B, type I receptors.

Nascent HDL interacts with the adenosine triphosphate binding cassette A1 (ABCA1) protein expressed on peripheral cells. ABCA1 facilitates active transport of cellular cholesterol to the nascent HDL. The nascent HDL is converted into a mature HDL via the lecithin: cholesterol acyltransferase (LCAT) enzyme. LCAT is a plasma enzyme that combines cholesterol and FFAs to form cholesteryl esters (Fig. 28.3). The mature HDL particle is larger and more spherical than the disk-shaped nascent HDL. This shape leads to a conformational change exposing the apoA1 binding site-specific for the scavenger receptor class B, type I (SR-BI). SR-BI is expressed on adrenal glands and hepatocytes and facilitates rapid cellular internalization of HDL. Mature HDL can also gain apoE, which enables additional hepatic clearance by LDLR and LRP. Once HDL is removed from circulation, the cholesterol it contained can be converted into hormones, repackaged as VLDL cholesterol, or excreted into the gastrointestinal tract as bile acids.

Neutral lipids (cholesteryl esters and triglycerides) can be shuttled between HDL and apoB-containing particles via the presence of a cholesteryl ester transfer protein (CETP). CETP is found primarily bound to HDL particles and allows for the equimolar exchange of lipids with other lipoproteins. Thus in patients with high LDL cholesterol, the concentration gradient of cholesterol to triglycerides results in a relatively cholesterol-rich HDL fraction (cholesteryl esters go into HDL while triglycerides go out). Conversely, in patients with high triglycerides carried on VLDL, the HDL particles become more triglyceride-rich. Individuals with homozygous mutations in *CETP* with no measurable CETP activity display high HDL-C, low LDL-C, and lower incidence of CVD. However, all attempts to date to create a pharmaceutical agent to inhibit CETP activity have failed to lower CVD risk substantially.

Clinical laboratory lipid measurements

The basic lipid panel

The basic lipid panel consists of three measured values: total cholesterol, HDL cholesterol, and triglycerides. Lipid panel measurements are almost exclusively performed on automated chemistry analyzers in a central laboratory. In addition, some point-of-care methods are also available that, while more expensive to perform, can aid in lipid management in select settings (health fairs, physician office clinics, or pediatric centers), where sending the patient to phlebotomy for central laboratory lipid testing is not ideal.

Total cholesterol

The method used to measure total cholesterol is an enzymatic reaction in which cholesteryl esters are first converted into unesterified cholesterol by cholesteryl ester hydrolase. This free cholesterol is then oxidized by cholesterol oxidase to cholest-4-en-3-one and hydrogen peroxide. Phenol-4-antipyrine and peroxidase then converts the hydrogen peroxide into a quinoneimine dye, which is measured spectrophotometrically (Fig. 28.4).

It is important to note that this assay does not distinguish between the esterified and the unesterified cholesterol. Approximately 60%–80% of cholesterol is esterified with FAs, largely as a result of the enzyme LCAT. The primary or secondary causes of LCAT deficiency (LCATD) can result in erythrocyte abnormalities and dramatically decreased esterification of serum cholesterol. Therefore when abnormalities in the cholesterol esterification process are suspected, a modified version of the cholesterol assay can be utilized, which omits the conversion of esterified cholesterol into free cholesterol to measure only the free cholesterol in circulation. The free cholesterol can then be compared with the total cholesterol in the same sample to determine the percentage of esterified cholesterol. A cholesteryl ester fraction of <20% indicates familial or acquired LCATD and increased risk of kidney, liver, and vascular diseases.

Triglycerides

Clinical laboratory methods to measure triglycerides are also enzymatic assays. Triglycerides are by nature heterogeneous in composition, containing FAs of various carbon-chain length and saturations. To circumvent this heterogeneity, clinical methods take advantage of the fact that all triglycerides include exactly one glycerol molecule. First, lipase is added to convert all triglycerides into free glycerol and FFAs. Then glycerol concentrations are quantified following a secondary reaction with glycerol kinase, glycerolphosphate oxidase, and a chromophore (Fig. 28.4).

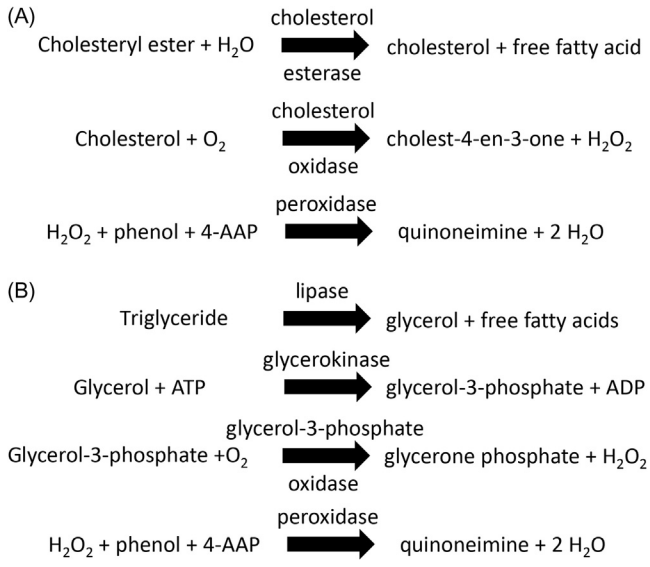


FIGURE 28.4 Enzymatic methods for measuring (A) cholesterol and (B) triglycerides.

The standard triglyceride assay used in most clinical laboratories will measure artificially high triglyceride concentrations (i.e., pseudohypertriglyceridemia) in patients that have abnormally elevated free glycerol. Increased plasma glycerol occurs in chronic alcohol drinkers, patients with glycerol kinase deficiency, or in cases of glycerol contamination of the specimen. If pseudohypertriglyceridemia is suspected, a glycerol-blanked assay can be used in which any free glycerol is first eliminated prior to the hydrolysis of the triglycerides, so that only triglycerides are measured. A comparison of the glycerol-blanked assay with the nonblanked assay can be used to determine the amount of free glycerol in circulation.

High-density lipoprotein cholesterol and nonhigh-density lipoprotein cholesterol

To measure HDL-C, most modern laboratories employ direct homogeneous assays, which utilize masking reagents that selectively interact with non-HDL lipoprotein particles and prevent cholesterol measurement. After masking, the standard enzymatic cholesterol reaction is performed to measure only the cholesterol within the particle of interest. Some reference laboratories and specialty lipid testing centers still employ traditional methodology for HDL-C measurement involving the precipitation of apoB-containing lipoproteins (LDL, IDL, VLDL, and chylomicrons) using a polyanion (such as dextran sulfate or heparin) and a divalent cation (such as calcium or manganese). After the non-HDL particles are precipitated, the cholesterol within the HDL is then measured enzymatically. This technique can still be useful in the cases of hypertriglyceridemia and diabetes where homogeneous methods tend to overestimate actual HDL-C [1–3]. However, new homogeneous methods are reportedly more reliable [4].

The 2001 National Cholesterol Education Program (NCEP) recommendations described a calculated parameter, non-HDL cholesterol, which should be included in basic lipid panels [5]. Non-HDL-cholesterol is calculated from a simple formula [see Eq. (28.i)]:

$$\text{Non-HDL-}C = \text{Total cholesterol} - \text{HDL-}C \qquad (28.\text{i})$$

Non-HDL cholesterol has been shown to predict future cardiovascular events. Non-HDL-C accounts for the cholesterol within all proatherogenic lipoprotein particles (LDL, lipoprotein (a) [Lp(a)], IDL, VLDL, and chylomicron remnants).

Low-density lipoprotein cholesterol

LDL-C is the focus of most lipid interventions and CVD risk assessments [6]. There are a variety of clinical LDL-C methods. We have chosen to focus on the most common clinical methods (estimation and direct homogeneous) and the gold-standard analytical method in this chapter.

Beta-quantification of low-density lipoprotein cholesterol

Beta-quantification is the gold standard method for LDL-C measurement. The assay name originates from early studies showing LDL particles migrate to the "beta" region on lipoprotein electrophoresis (see subsequent section on electrophoresis methods). The contemporary beta-quantification reference method was adopted by the NCEP in 1995 and uses a single ultracentrifugation step at $120{,}000 \times g$ for 16 hours to separate particles at 1.006 g/mL in which chylomicron and VLDL particles will float to the top layer of the sample (Fig. 28.5). The top layer is removed from the bottom layer, which contains LDL, Lp(a), HDL, and IDL [7,8]. Cholesterol is measured in this bottom fraction before it is treated with a divalent cation to precipitate any apoB-containing particles [i.e., LDL, Lp(a), and IDL]. After the precipitation, HDL is the only particle left in solution. Cholesterol is then measured in the HDL fraction. The HDL cholesterol is subtracted from the cholesterol measured in the LDL, Lp(a), HDL, and IDL layer to obtain the cholesterol in LDL (Fig. 28.5).

Prior to standardization to NCEP beta-quantification, LDL-C was classically defined as the measure of all cholesterol with a density between 1.019 and 1.063. Beta-quantification effectively altered the definition to cholesterol from *apoB-containing* lipoproteins with a density ≥ 1.006 mg/dL [8,9]. The amount of cholesterol in IDL or Lp(a) in most subjects is substantially less than the amount of cholesterol in LDL, so they are excluded in the measurement of LDL-C by beta-quantification. However, in the cases of dyslipidemias in which either Lp(a) or IDL is substantially increased, the measured concentration of LDL-C by beta-quantification will be artificially elevated. This inclusion is likely beneficial to the clinical performance of contemporary LDL cholesterol as a biomarker, due to the atherogenic nature of both IDL and Lp(a) [10]. However, it emphasizes the heterogeneity and lack of molecular specificity in the measure of LDL-C.

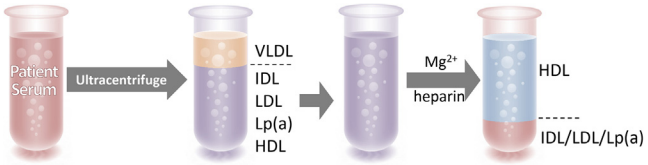


FIGURE 28.5 Beta-quantification. Patient sera is ultracentrifuged and the fraction with density <1.006 is discarded. The cholesterol content in the remaining fraction is measured before and after apolipoprotein B lipoproteins are precipitated.

Estimation of low-density lipoprotein cholesterol

The most commonly used clinical LDL-C method is not a measurement at all, but instead an estimation using the Friedewald equation [11]. In this calculation, the cholesterol within all lipoprotein particles except for LDL is subtracted from the total cholesterol. The main cholesterol-containing particles in a fasting sample are LDL, HDL, and VLDL. Total cholesterol and HDL-C are measured as described above, while VLDL cholesterol is estimated using the calculation: triglycerides (in mg/dL) divided by 5. Thus using the three measurements from a basic lipid panel, the Friedewald equation enables an estimate of the concentration of LDL cholesterol [see Eq. (28.ii)].

$$\text{LDL-}C = [\text{Total cholesterol}] - [\text{HDL-}C] - ([\text{Triglycerides}]/5) \tag{28.ii}$$

It should be recognized that the Friedewald equation was developed in a relatively small number of subjects and intended for research (not clinical) purposes. Nevertheless, estimating the cholesterol within LDL is much easier than measuring by beta-quantification, so the Friedewald equation was ubiquitously adopted by clinical laboratories and reported along with total cholesterol, HDL-C, and triglycerides. In Friedewald's original publication, the authors list three restrictions to the validity of estimated LDL cholesterol: (1) subjects must be fasting; (2) triglycerides must be <400 mg/dL; and (3) patients with a type III hyperlipidemia cannot be assessed [11]. Although the Friedewald estimation was developed for research, without it, LDL cholesterol would not have been incorporated so rapidly into standard clinical care [7].

Estimated LDL cholesterol relies on the assumption that triglycerides are exclusively within VLDL and that the VLDL triglyceride-to-cholesterol ratio is 5:1 (in mg/dL). The fasting requirement eliminates triglyceride contributions from chylomicrons, which typically have a much higher triglyceride-to-cholesterol ratio. The triglyceride limit of 400 mg/dL is meant to reduce inaccuracy of estimated LDL-C caused by the variability in VLDL triglyceride-to-cholesterol ratio at higher triglyceride concentrations. A more contemporary analysis of >1.3 million fasting patient samples with triglycerides <400 mg/dL reported a median VLDL triglyceride-to-cholesterol ratio of 5.2 (IQR 4.7–6.0) [12]. Moreover, the range of VLDL triglyceride-to-cholesterol ratio spanned 0.4–145 and was dependent on triglyceride, total cholesterol, and

HDL-C concentrations. Several studies have confirmed that Friedewald estimated LDL cholesterol is progressively less accurate with increasing triglycerides, with many studies suggesting a limit of 200 mg/dL triglycerides for accurate estimation of LDL-C [13,14]. Estimated LDL-C is also known to be less accurate in conditions associated with altered composition of triglycerides within lipoproteins such as diabetes mellitus [15] and type III hyperlipidemia [16]. In addition, in patients with very low LDL-C concentrations (i.e., below 70 mg/dL), the Friedewald equation tends to underestimate LDL-C [17,18]. These low LDL-C concentrations are becoming increasingly more prevalent as a result of more aggressive lipid-lowering therapies in randomized clinical trials showing increased clinical benefit. Therefore assuming a triglyceride-to-cholesterol ratio of 5:1 to all individuals can result in some inaccurate LDL-C estimations compared with LDL-C measured by beta quantification. However, these clinical scenarios may not be apparent at time of testing, and thus the inaccuracy of estimated LDL-C in these contexts often goes unnoticed.

Several iterations of improved LDL-C estimating equations have been developed, since the Friedewald equation was first proposed. The Martin–Hopkins equation is similar to the Friedewald equation but substitutes a variable in the denominator for estimating the VLDL cholesterol [see Eq. (28.iii)].

$$\text{LDL-}C = [\text{Total cholesterol}] - [\text{HDL-}C] - ([\text{Triglycerides}]/X) \quad (28.\text{iii})$$

The variable X changes according to an empirically derived table that adjusts for triglyceride values up to 400 mg/dL and non-HDL-C between 100 and 220 mg/dL. Using many different VLDL-C factors provides slightly more accurate LDL-C estimations [12], but does not fully correct for the inaccuracies attributable to estimated LDL-C [19]. Nevertheless, some clinical laboratories have acquired licenses to use this proprietary equation.

The empirically defined table on which the Martin–Hopkins equation is based can be mathematically fit using a bivariate quadratic equation with non-HDL-C and triglycerides as inputs. A more generalizable estimate was developed by the National Institutes of Health (NIH) using this approach [see Eq. (28.iv)] [20].

$$\text{LDL-}C = \frac{\text{TC}}{0.948} - \frac{\text{HDL-}C}{0.971} - \left(\frac{\text{Trig}}{8.56} + \frac{\text{Trig} \times \text{non-HDL-}C}{2140} - \frac{\text{Trig}^2}{16,100}\right) - 9.44 \quad (28.\text{iv})$$

The use of continuous variables rather than a delimited table provides the NIH equation superior comparability to measured LDL-C methods at triglyceride concentrations up to 2000 mg/dL and at LDL-C $<$70 mg/dL.

It should be noted that historically all calculated LDL-C was derived using the Friedewald equation. However, given its limitations, some laboratories have implemented other equations when reporting calculated LDL-C, whether it is the Hopkins/Martin equation (which is proprietary and requires a license to use) or the NIH equation. Therefore estimated LDL-C results from different institutions can no longer be assumed to be the same.

Direct homogeneous low-density lipoprotein cholesterol

Direct homogeneous LDL cholesterol assays performed on autoanalyzers are also available and require no special sample preparation or precipitation. Various methods are employed by the different manufacturers, relying on detergents or modified enzymes to isolate selectively and measure only the cholesterol within the LDL particles [21]. These assays are easy to use, cost-effective, and cleared by the US Food and Drug Administration (FDA). However, they suffer from inaccuracies, especially in the cases of dyslipidemias. Bias between 13% and 46% has been reported by various FDA-cleared direct LDL methods [7,22].

Lipoprotein particles of all classes are heterogeneous in size, density, lipid content, and protein composition. When patients have abnormal lipid profiles, the normal distribution of these lipoprotein subclasses is altered. Samples with atypical lipoprotein distributions tend to give inaccurate results when measured by direct LDL-C methods. Various direct LDL-C methods have differing specificities for lipoproteins and measure up to 24% of VLDL cholesterol, 72% of IDL cholesterol, 90% of Lp(a) cholesterol, and even 18% of HDL cholesterol [23,24]. Therefore these assays are only suitable to measure LDL-C in normal patients without dyslipidemia. Laboratorians need to be aware of these limitations and caution clinician colleagues regarding these weaknesses.

Lipoprotein electrophoresis

Lipoprotein electrophoresis is a method that can be used in special cases where dyslipidemias are suspected [25]. In this method, a patient serum is applied to agarose gel and a current is applied allowing for negatively charged lipoproteins to migrate toward the anode. The combination of size and charge of each lipoprotein will determine its migration from the site of sample application. A nonspecific neutral lipid stain such as Sudan Black can then be applied to visualize lipoprotein bands (Fig. 28.6). More specific enzymatic cholesterol staining methods enable quantification by scanning densitometry. In this method, the HDL particles travel the farthest, since they are small and have many negatively charged apolipoproteins and migrate to what is called the “alpha” band. In contrast, LDL does not travel as far as HDL, because it is slightly larger than HDL and has fewer surface proteins.

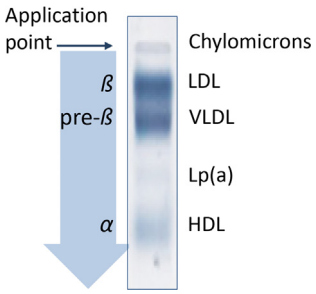


FIGURE 28.6 Lipoprotein electrophoresis. The banding pattern shown is from a serum pool containing chylomicrons, low-density lipoprotein, very-low-density lipoprotein, lipoprotein(a) and high-density lipoprotein.

The position in which LDL resides on electrophoresis is called the "beta" band. VLDL migrates between LDL and HDL in a region called the prebeta band. Even though it is larger than LDL, VLDL has many more surface apolipoproteins, which drives its migration closer to the anode. Patient samples with elevated IDL will often show bridging between the beta and prebeta bands, resulting in a broad beta band. Chylomicrons are so large and neutrally charged that they do not migrate out of the point-of-sample application. Lipoprotein electrophoresis is useful in the cases of suspected dyslipidemia where more traditional lipid testing is inadequate to resolve the diagnosis. It is also the best method to detect the abnormal lipoprotein Lipoprotein X (LpX), a unique lipoprotein found in the cases of cholestasis that reverse migrates (i.e., toward the cathode; see Section "Lipoprotein X").

Apolipoproteins

Other methods can be used to measure directly the amount of apolipoproteins by automated immunoassays on autoanalyzers such as apoB [contained in LDL, Lp(a), IDL, and VLDL], apoA-1 (containing HDL), and apolipoprotein (a) [apo(a); contained in Lp(a)]. All of these methods are becoming more routine, since they are compatible with nearly all central laboratory analyzers as either immunoturbidimetric or immunonephelometric assays.

Apolipoprotein B

ApoB has been proposed by several lipid and cardiology experts as an alternative to LDL cholesterol for the assessment of cardiovascular risk and diagnosis of dyslipidemias [26–28]. Unlike LDL-C, which is confused by the variety of methods with varying specificities and definitions, apoB is a well-characterized protein. A standardized reference material for apoB is available from the World Health Organization in collaboration with the International Federation of Clinical Chemistry [29]. Comparison of 22 different clinical apoB methods found a bias <5% between all methods and 3.6% with imprecision <3.2% [30]. Consequently, apoB methods are significantly more accurate and precise than LDL cholesterol [7].

Apolipoprotein A-I

ApoA1 is useful in the cases of suspected HDL deficiencies. As described above, apoA1 is responsible for reverse cholesterol transport, which reduces atherosclerotic lesions in arteries. Thus the apoB:ApoA1 ratio provides insight into the balance between the atherogenic and antiatherogenic lipoproteins. Several large prospective studies have shown the apoB:apoA1 ratio to be a superior predictor of CVD risk compared with LDL-C [31–33].

Apolipoprotein (a) Apo(a) is the unique apolipoprotein found in Lp(a). Immunoassays that measure Lp(a) utilize antibodies that recognize the apo(a) protein and are useful for the identification of individuals that express high concentrations of this extremely proatherogenic particle. It is especially important in individuals with a strong personal or family history of early CVD events without elevated LDL-C or in individuals where lipid-lowering medications do not reduce their LDL-C sufficiently [remember, the cholesterol within Lp(a) is included in nearly all methods for LDL-C determination; see Section "Lipoprotein (a)"].

Nonfasting lipid assessment

Clinical practice guidelines from major cardiology societies and consensus statements have concluded that nonfasting blood samples are acceptable for routine lipid assessments [34–39]. Fasting samples were historically required for lipid assessment in order to reduce the variability observed in measures of triglycerides (and consequently estimations of LDL-C) [11]. Several lines of evidence support the shift to nonfasting lipid assessment. Data from large groups of patients show that observed triglyceride increases due to nonfasting are clinically negligible in the vast majority of cases. Several large studies have confirmed that nonfasting triglyceride concentrations remain <200 mg/dL in a majority of patients [40–42]. This relatively normal concentration of triglycerides in the nonfasting state contributes to the insignificant impact observed by nonfasting lipid screening and LDL-C estimation [41,43]. The high biological variability of triglycerides is another argument in favor of nonfasting lipid assessment. The average biological variability of *fasting* triglycerides is 20%–30%, and ranges from <5% for persons with low fasting triglycerides (<100 mg/dL) to as large as 75% for persons with elevated fasting triglycerides [44]. Consequently, postprandial increases are often within the noise of typical biological variation [45].

Due to the long history of fasting lipid assessment, it would be prudent to report nonfasting triglycerides and estimated LDL-C unique from fasting samples. In this way, nonfasting triglycerides can have a unique abnormal flag at the recommended 200 mg/dL cut point and an automatically appended comment suggesting repeat testing after an 8−12-hour fast. Due to the novelty of reporting nonfasting Friedewald estimated LDL-C, an appended comment suggesting values may be decreased when triglycerides >200 mg/dL may also be reasonable.

Lipoprotein subfraction methods

Many clinical methods that measure lipoprotein subfractions are available (e.g., nuclear magnetic resonance, vertical gradient ultracentrifugation, ion-mobility, gradient gel electrophoresis, etc.). These methods distinguish between different sizes and densities of LDL and HDL with increased granularity. However, the methods separate by different means (i.e., size, density, and charge), define small, medium, or large by different limits, and measure different lipoprotein components [46]. No medical society has endorsed the use of lipoprotein subfractions. None of the subfraction methods in clinical use have sought FDA clearance for subfractionation of lipids, but rather emphasized their substantial equivalence to traditional lipid measures. The most recent LDL subfractionation method to be cleared by the FDA included an outcome study looking at association between LDL subfraction and CVD in the Atherosclerosis Risk in Communities cohort [47]. In the published report, the authors concede that "small dense LDL-C was not significantly associated with risk for incident coronary heart disease after further adjustment for other lipid risk factors, such as LDL-C, apo B, or total cholesterol."

The lack of independent predictive information for lipoprotein subfractions has been repeatedly cited in many opinion papers and consensus documents [48−50]. The National Lipid Association, the National Academy of Clinical Biochemistry, the American College of Cardiology (ACC), and the American Heart Association (AHA) have all published guidelines specifically recommending *against* the use of lipoprotein subfraction methods for initial clinical assessment or on-treatment management decisions [51−53].

Lipoprotein disorders

Lipid disorders were originally classified by Fredrickson in 1965 into different types (historically termed Fredrickson phenotypes) [54]. The phenotypes were based on patterns of lipoprotein concentrations, as shown by lipoprotein staining on agarose gel electrophoresis. Five of the six classic Fredrickson phenotypes include elevations of triglycerides [55,56]. However, at the time of classification, the underlying disorders were not well characterized. Today, we know that phenotypes can have multiple etiologies from either primary and/or secondary causes, and several lipid diseases do not fit into a Fredrickson phenotype. Both classic phenotypes and other lipoproteinemias are listed in Table 28.1 with their expected laboratory results and major etiologies.

Hyperlipoproteinemias

Exogenous hyperlipemia

The type I Fredrickson phenotype of extremely elevated chylomicrons is associated with extremely elevated plasma triglycerides usually greater than 1000 mg/dL in the fasting state. This rare phenotype, called familial chylomicronemia syndrome (FCS), occurs in 1 in 100,000 to 1 in 1,000,000 individuals and is most commonly caused by the absence of lipoprotein lipase (LPL) activity from the genetic mutations of both LPL alleles. Less commonly, FCS can be caused by mutations in other genes required for proper LPL localization or function, such as the loss of the cofactor apolipoprotein C-II, which is required for LPL activity [57]. LPL is required to hydrolyze triglycerides and distribute FAs to metabolically active tissues; therefore patients with type I hyperlipoproteinemias form no chylomicron remnants; intact chylomicrons stay elevated in systemic circulation, and triglycerides are shuttled to HDL particles via CETP. Triglycerides-rich HDL is rapidly removed from circulation resulting in very low plasma HDL-C concentrations. The loss of remnant lipid flux to the liver results in decreased production of apoB-containing lipoprotein and low plasma VLDL-C and LDL-C concentrations. Chylomicrons are too large to invade the vascular wall so atherosclerosis is not usually a concern.

Depending on dietary fat intake, abdominal pain and pancreatitis are the main clinical consequences of extremely elevated triglycerides via a mechanism that has not been fully elucidated [58]. The resulting hypertriglyceride-induced pancreatitis is more severe with worse outcomes when compared with pancreatitis caused by other etiologies [59]. Children with FCS are identified by extremely elevated triglyceride concentrations or failure to thrive and further clinical sequelae such as eruptive xanthomas, lipemia retinalis, hepatosplenomegaly, or a lifetime risk of acute pancreatitis often manifest by adolescence. Lipid testing reveals a triglyceride-to-cholesterol ratio that is similar to what is found in chylomicrons (>10:1), and lipoprotein lipase activity is decreased. Chylomicrons can also be detected electrophoretically. They are extremely large and do not migrate well in gel electrophoresis systems, so they are visualized as a band at the sample application point. With no

TABLE 28.1 Lipoprotein disorders.

Dyslipidemia	Fredrickson phenotype	Abnormal lipoprotein(s)	Cholesterol (mg/dL)	Triglycerides (mg/dL)	Frequency	Primary disorder(s)
Exogenous hyperlipemia	Type I	↑Chylomicrons	Any	>10,000	1:1,000,000	Lipoprotein lipase deficiency; ApoC-II deficiency
Hypercholesterolemia	Type IIa	↑LDL	>200	<250	1:500	Familial hypercholesterolemia
Combined hyperlipidemia	Type IIb	↑LDL & VLDL	>200	>250	1:100	Familial multiple lipoprotein-type hyperlipidemia
Remnant hyperlipidemia	Type III	↑IDL (a.k.a. ß-VLDL)	Any	Any	1:500	Familial dysbetalipoproteinemia
Endogenous hyperlipemia	Type IV	↑VLDL	Any	>250	1:100	Familial hypertriglyceridemia (mild)
Mixed hyperlipemia	Type V	↑Chylomicrons ↑VLDL	Any	>1000	1:600	Familial hypertriglyceridemia (severe); Familial lipoprotein lipase deficiency
Lipoprotein (a)	–	Lp(a)	Any	Any	1:5	Elevated Lp(a)
Lipoprotein X	–	LpX	>250 LDL-C >200	Any	1:5 among low HDL-C (<10 mg/dL)	Cholestatic liver disease or LCAT deficiency
Hypobetalipoproteinemia	–	↓LDL	<100	<50	1:3000	Familial hypobetalipoproteinemia
Abetalipoproteinemia	–	↓↓ LDL	Undetectable	Undetectable	<1:1,000,000	Microsomal triglyceride transfer protein deficiency
Hypoalphalipoproteinemia	–	↓↓HDL	Any	Any	<1:100	Familial ApoA1 deficiency; ABCA1 deficiency (Tangier Disease); LCAT deficiency

ABCA1, Adenosine triphosphate binding cassette A1; *ApoA1*, apolipoprotein A-I; *ApoC-II*, apolipoprotein C-II; *HDL-C*, high-density lipoprotein cholesterol; *IDL*, intermediate-density lipoproteins; *LCAT*, lecithin: cholesterol acyltransferase; *LDL*, low-density lipoprotein; *LDL-C*, low-density lipoprotein cholesterol; *Lp(a)*, lipoprotein(a); *LpX*, lipoprotein X; *VLDL*, very-low-density lipoproteins.

FDA-approved treatments, patients with FCS must adhere to a stringent diet that is nearly completely devoid of fat. Even with extreme caution toward food intake, triglycerides rarely reduce to normal concentrations in patients with FCS. A homozygous or compound heterozygous genetic variant is sometimes absent and secondary causes of hypertriglyceridemia (high-fat diet, alcohol consumption, metabolic syndrome, type 2 diabetes, hypothyroidism, or certain medications) should be considered.

Familial hypercholesterolemia

Type IIa, Familial hypercholesterolemia (FH), is the only Fredrickson phenotype that does not have elevated plasma triglyceride concentrations. It is an autosomal codominant disorder characterized by elevated cholesterol, which is predominantly found in LDL. FH is further subcategorized as either heterozygous or homozygous, which phenotypically differ in the magnitude of LDL-C increase as well as in resulting risk for developing CVD. Individuals with homozygous FH (hoFH) have significantly higher LDL-C (500−1200 mg/dL) than heterozygotes FH (hetFH), who typically have LDL-C concentrations of 190 mg/dL or greater. Early onset CVD occurs in hetFH individuals in the fifth or sixth decade of life, as compared with childhood in hoFH individuals. Variants in one or both alleles that inactivate either the LDLR or apoB, or mutations that cause a gain of function of proprotein convertase subtilisin/kexin type 9 (PCSK9), can cause FH. The liver requires the action of the LDLR to take up LDL particles from circulation and dispose of the cholesterol contents. apoB is the binding partner for LDLR, so mutations in either apoB or LDLR result in a dramatic increase of LDL cholesterol resulting in FH. PCSK9 functions to remove LDLR from hepatocytes, so gain of function mutations in PCSK9 will cause the FH phenotype. The vast majority of FH cases ($>80\%$) are a result of mutations in LDLR, 5%−10% are from apoB mutations, and $<1\%$ are from PCSK9 gain of function mutations.

The prevalence of heterozygous FH is now estimated to be around 1:250 [60], with some specific populations having a higher prevalence of 1 in 50−67 [61]. Homozygous FH is less common, with a prevalence of 1 in 300,000 [62]. Three groups have developed diagnostic tools to identify individuals with FH (the US MedPed Program, the Simon Broome Register Group in the United Kingdom, and the Dutch Lipid Clinic Network). In general, individuals with FH can be identified by the following criteria: an elevated LDL-C on a standard lipid screen, a family history of hypercholesterolemia and/or early CVD events, personal premature CVD, or the presence of tendon xanthomas. The more the criteria that are met for a patient, the higher the likelihood of FH. Patients with FH will have elevated blood cholesterol starting in childhood. It is now recognized that the lifetime arterial exposure to elevated cholesterol is directly correlated with future risk of developing CVD (similar to the concept of increased lung cancer risk with increased "pack years" of smoking). This means that people with genetic FH have been exposed to high cholesterol from birth and have higher risk of CVD as compared with individuals that may acquire hyperlipidemia later in life from poor diet or lifestyle. Therefore identifying individuals with FH as early in life as possible is important to minimize their risk of future CVD development via early lifestyle and medication therapies.

Combined hyperlipidemia

The Fredrickson-type IIB phenotype is characterized by elevated LDL and VLDL with corresponding elevations in both cholesterol and triglycerides. Also referred to as familial combined hyperlipidemia (FCH or FCHL), the phenotype is caused by hepatic overproduction and/or decreased clearance of apoB-containing lipoproteins (VLDL, IDL, and LDL) and apoB concentrations exceed the 90th percentile of a reference population [63,64]. It is the most prevalent primary dyslipidemia, with a prevalence of 1%−3% in the general population; however, a lack of primary care physician awareness often leads to misdiagnosis. Susceptibility to FCHL is higher in certain ethnic groups such as Hispanics (4.8% prevalence) and is commonly present along with other metabolic diseases such as obesity, insulin resistance, metabolic syndrome, and type 2 diabetes mellitus. Individuals with FCHL have a higher risk of premature coronary artery disease, because both the cholesterol-rich lipoproteins and triglyceride-rich lipoproteins that present in this phenotype can cause atherosclerosis. FCHL is generally considered an oligogenic disease such that genetic diagnosis is not yet possible, but the finding of mixed hyperlipidemia, elevated apoB, and a family history of mixed hyperlipidemia strongly suggests the FCHL phenotype. No specific therapies have been developed for FCHL, so treatment usually aims to control modifiable cardiovascular risk factors such as obesity, smoking, alcohol intake, and cholesterol concentration. In addition, consideration is given to avoid secondary causes of hypertriglyceridemia or implementing pharmacological triglyceride lowering therapies.

Remnant hyperlipidemia

The type III phenotype, or remnant hyperlipoproteinemia, is characterized by elevated remnant particles of triglyceride-rich lipoproteins—namely chylomicron remnants and VLDL remnants. This phenotype is mainly caused from an inability of the liver to clear these remnant particles from circulation [65]. Rarely, individuals

will have a genetic mutation that causes significantly reduced expression of the apoE protein that is required for hepatic clearance of IDL and other RLPs. However, more commonly, these individuals are homozygous for the apoE2 allele that has dramatically decreased affinity for LDLR and LRP (termed familial dysbetalipoproteinemia). The circulating remnants can pick up cholesterol from other lipoproteins such as HDL via lipid exchange mediated by CETP. Therefore accumulated IDL and remnant particles contain both cholesterol and triglycerides, so both are significantly elevated in plasma [66]. In addition to elevated triglycerides and cholesterol concentrations, these patients clinically present with lipid deposits in their skin such as striated palmar xanthomas or tuberous xanthomas. On lipoprotein electrophoresis methods, LDL runs at the β band and VLDL runs at the pre-β band. The abnormal lipoproteins present in type III dysbetalipoproteinemia (IDL and VLDL remnants) migrate between the LDL and VLDL bands in a "broad beta band." These particles are sometimes referred to as β-VLDL particles (i.e., VLDL that migrate in the "beta" band) [67]. Remnant particles are extremely atherogenic, so patients are at high risk for atherosclerosis and CVD. ApoE2 homozygotes are fairly common (~1% prevalence) but dysbetalipoproteinemia occurs in <5% of these individuals, so other contributing factors, possibly a concomitant FCH phenotype, are needed to cause the phenotype.

Endogenous hyperlipemia

Type IV hyperlipoproteinemia is characterized by increased VLDL and triglycerides and affects approximately 1 in 100 individuals. The condition can be genetic and passed down through families as familial hypertriglyceridemia. The condition can be caused by an increased hepatic production of VLDL or from heterozygous LPL deficiency. The majority of cases are caused by multigenetic effects and/or secondary causes of hypertriglyceridemia [68]. Triglyceride concentrations are elevated, but are usually lower (<1000 mg/dL) than that observed in the more severe type I phenotype, because VLDL contains less triglycerides as compared with chylomicrons. However, if high enough, triglycerides can still cause pancreatitis. Total cholesterol and LDL-C concentrations are usually within normal target concentrations. Individuals with this condition tend to have low HDL-C, metabolic syndrome, insulin resistance, and obesity and have precocious atherosclerosis and increased risk of CVD events. Treatment starts with lifestyle modifications such as increasing exercise while lowering carbohydrate and alcohol ingestion. Pharmacological treatments such as prescription omega-3 FAs, fibrates, or niacin can be considered.

Mixed hyperlipemia

Type V hyperlipidemia is characterized by high triglycerides in both chylomicrons and VLDL. This hypertriglyceridemia is either the heterozygous form of the mutations that cause type I FCS (e.g., heterozygous LPL deficiency) or can be polygenic in origin with familial clustering and a prevalence of approximately 1 in 600 individuals. Similar to FCS, patients with polygenic type V hyperlipidemia have extremely elevated triglycerides >1000 mg/dL when triggered by a secondary cause, milky plasma, and a risk of pancreatitis. Since proatherogenic VLDL particles are also increased in this phenotype, these patients are susceptible to increased risk of CVD. In addition, they respond well to triglyceride lowering therapies such as fibrates, niacin, or fish oil.

Distinguishing between the hypertriglyceridemia disorders can be made using advanced lipid testing such as lipoprotein electrophoresis in combination with standard biochemical lipid measurements. A qualitative assessment can also be made visually given that after a sample with high triglycerides is allowed to sit refrigerated overnight. If the sample contains both chylomicrons and VLDL (i.e., a type V), the chylomicrons will float to the top producing a cream-like supernatant, while the elevated VLDL will produce a cloudy lower fraction. If the sample contains only chylomicrons (i.e., a type I), the cream-like supernatant will appear while the lower fraction will be clear. If only VLDL is elevated (i.e., a type IV), the sample will be cloudy throughout and no cream layer will form.

Lipoprotein (a) disorder

Lp(a) (pronounced "lipoprotein little 'a'") is a unique and highly atherogenic lipoprotein particle (Fig. 28.7) [69]. At its core, Lp(a) is an LDL-like particle; however, Lp(a) has a unique feature in that the apoB protein is covalently bound to a unique and highly glycosylated protein called apo(a) (pronounced "apo-little a"). Apo(a) shares sequence homology with, and has likely evolved from, the proenzyme plasminogen. While plasminogen contains five kringle domains (KI−KV) and a protease domain, apo(a) has an inactive protease and does not express the first three kringle domains but does contain 10 subtypes of KIV domain. The second KIV (KIV_2) is present in variable ($n = 1$ to >40) copies. This variability of kringle length is genetically defined and causes the size of the Lp(a) particle to vary greatly with most individuals expressing two different sized apo(a) isoforms, one from each parent. Concentrations of Lp(a) in circulation are primarily determined by the *LPA* gene locus and are not influenced by environment or dietary factors. The half-life of Lp(a) in circulation seems to be longer than LDL,

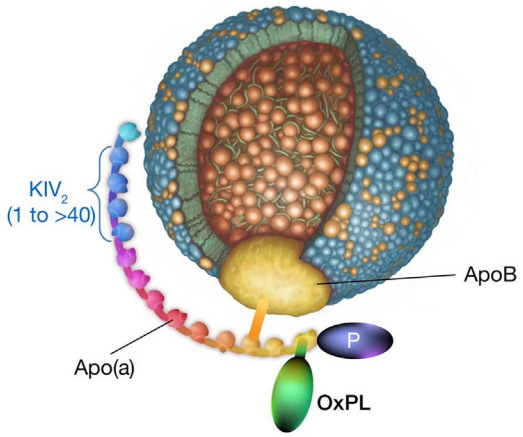


FIGURE 28.7 Lipoprotein(a) is a potently atherogenic lipoprotein. Lipoprotein(a) particles are comprised of an low-density lipoprotein-like core with the apolipoprotein(a) covalently bound to apolipoprotein B by a disulfide bridge. Apolipoprotein(a) contains domains called kringle domains with a variable number of second KIV repeats (shown here as four repeats). Oxidized phospholipids also interact with apolipoprotein (a).

possibly due to apo(a) interference of apoB binding to LDLR for LDL metabolism. In addition, there is an inverse correlation between apo(a) size and Lp(a) concentration likely because the liver can synthesize smaller apo (a) proteins faster than larger ones. The mechanism(s) of Lp(a) metabolism and clearance are not fully elucidated.

Observational studies, prospective population studies, Mendelian randomization studies, and genome-wide association studies have all shown that Lp(a) is highly atherogenic. It contributes to both CVD and calcific aortic valve stenosis [70]. Similar to LDL, Lp(a) particles carry cholesterol and can enter vessel walls, oxidize, and create immunogenic and proinflammatory molecules that promote atherosclerosis. However, apo(a) also contributes to the atherothrombosis via potential plasminogen-like antifibrinolytic effects and interactions with oxidized phospholipids at the KV subunit of apo(a), among others. Given the increased thrombosis at sites of plaque rupture, apo(a) has the potential to cause myocardial infarctions and ischemic stroke. The risk for myocardial infarction, stroke, and peripheral arterial disease associated with Lp (a) expression has been shown in many studies with significant risk starting when Lp(a) measures 30 mg/dL or higher using an automated immunoassay [71]. However, several guidelines, including the most recent 2018 lipid management guidelines by the AHA/ACC and a 2019 position statement by the National Lipid Association, direct physicians to consider intensifying CV risk reduction strategies when Lp(a) is >50 mg/dL. It is estimated that around 20% of the population has elevated Lp(a). However, different ethnic groups display different median Lp(a) concentrations (median/interquartile range: Caucasians: 12/5−32 mg/dL; Hispanics: 19/8−43 mg/dL; Black: 39/19−69 mg/dL) [72]. Whether non-Caucasian should use the same risk cut points as Caucasians is unknown so universal cut points are currently used for test interpretation.

Measuring Lp(a) is not straightforward given that the size of Lp(a) differs depending on the size of the apo(a) isoform(s) expressed in each patient. Commercially available immunoassays may utilize polyclonal antibodies that can cross-react with the KIV_2 region, so clinical laboratories should use assays that are the least susceptible to Lp (a) isoform size [73]. Historically, nearly all studies have used assays that report Lp(a) concentration in mass units of mg/dL. However, the differences in isoform size within the population can lead to inaccurate results when using an assay that is calibrated to mass units. The resulting bias can lead to underestimation of concentration in small Lp(a) isoforms and overestimation in large Lp(a) isoforms; thus there is potential for miscategorizing patients with Lp(a) concentrations near clinically defined cut points. The use of appropriate calibrators using a 5-point calibration scheme and an effort to standardize results to the World Health Organization/International Federation of Clinical Chemistry and Laboratory Medicine International Reference Reagent has helped to minimize the influence of isoform size on measured concentration [73]. Calibrating assays to molar units of nanomoles per liter (nmol/L) alleviates the inaccuracies caused by calibrators of different molecular weights. Indeed, the international reference material is assigned molar concentration units. While some assays are traceable to these reference calibrators, others are not, and results from different commercially available assays are not harmonized. It has been recommended that laboratories cease reporting of Lp(a) in mass units (mg/dL) and instead report in the more accurate units of particle number (nmol/L). However, making this change may be challenging for several reasons. Currently, assay manufacturers may not provide reagents calibrated to particle number or those reagents are not FDA-cleared (an important consideration in the United States) using this calibration scheme. Given that different Lp(a) isoforms have different molecular weights, results from assays calibrated to the traditional mg/dL units cannot be accurately converted into nmol/L after testing. Most clinical studies have reported Lp(a) results in units of mg/dL and most guidelines still use these units, so physician education may be challenging when reporting units are changed. In addition, clinical cut points using particle numbers have not been agreed upon. For example, the 2018 ACC/AHA cholesterol guidelines state that a value of >125 nmol/L (or >50 mg/dL) should be used to consider a patient at higher risk from Lp(a) expression; however, the 2019 National Lipid Association position statement recommends using >100 nmol/L [69].

Testing for Lp(a) is generally recommended when the patients have a personal of family history of premature CVD or if they are at intermediate risk and therapy decisions are uncertain, however some advocate for measuring it once in everyone. Since expression is mainly genetically determined and not affected greatly by lifestyle of pharmaceutical treatment, repeat measurement is not recommended. There is currently no treatment for Lp(a) that specifically targets Lp(a) particles separately from other lipoproteins and concomitantly lowers CV risk. Niacin has been shown to lower Lp(a), but side effects of treatment along with the absence of outcome benefit have led this treatment to fall out of favor. Mipomersen (no longer being produced) and PCSK9 inhibitors have been shown to lower Lp(a), but are only indicated in a very selected patient population. Lipoprotein apheresis removes Lp(a) from circulation and small studies have shown a benefit to outcomes, but this therapy is expensive, time-consuming, and not recommended for most individuals. Interestingly, there is some evidence that statin therapy may moderately increase Lp(a) expression when looking at individual patient Lp(a) values rather than population averages. Correspondingly, while statin therapy lowered the risk for CVD events in patients with elevated Lp(a), the risk lowering was blunted in comparison with individuals with lower Lp(a) [74]. Given the lack of effective targeted treatment for increased Lp(a), physicians will mainly attempt to lower overall CVD risk via intensifying standard cholesterol-lowering strategies. Emerging therapies using antisense oligonucleotide treatment that directly lower apo(a) expression and, correspondingly, Lp(a) concentration may hold significant promise.

Lipoprotein X

LpX is an abnormal vesicle produced during cholestasis and in patients with LCATD. LpX contains bile salts, albumin, a high proportion of unesterified cholesterol, and a phospholipid bilayer. LpX has a similar density to LDL, making LpX cholesterol indistinguishable from LDL-C by beta-quantification, direct homogeneous, and estimated LDL-C methods. Due to this similarity in densities, LpX is often responsible for false-elevations in LDL-C, and LpX should be suspected when unusually rapid elevations in LDL-C are encountered with a low apoB concentration. LpX is an indicator of acute liver dysfunction, which is often a more time-sensitive concern than CVD risk. LpX is not measured via standard lipid or lipoprotein analysis. It is only identified using lipoprotein electrophoresis, where LpX runs in a reversely migrating band. Staining with the standard neutral lipid stain (e.g., Sudan black or similar) is less sensitive for the detection of LpX and should only be used qualitatively. Use of filipin after electrophoresis has been recently developed as a more sensitive and quantitative detection method [75]. It should be noted, however, that LpX is a labile particle and unstable postsample freezing. When LpX is present, LDL-C should not be reported. Patients with LpX will often have dramatically reduced HDL-C concentrations.

Hypolipoproteinemias

While low cholesterol is considered protective against development of CVD, some disorders cause pathologically low concentrations of lipoproteins. Lipid concentrations can be low from secondary effects of conditions such as malnutrition, malabsorption, or chronic liver disease. However, several rare genetic conditions can also result in low concentrations of specific lipoprotein classes.

Hypobetalipoproteinemia

Most clinicians worry about elevated LDL-C; however, in rare cases, individuals can have a condition resulting in extremely low LDL-C (below the 5th percentile of the general population) and reduced apoB [76]. Familial hypobetalipoproteinemia type 1 (or homozygous hypobetalipoproteinemia) results from mutations in both *APOB* genes that produce variably truncated ApoB protein. The consequence is reduced by the synthesis of VLDL particles from the liver with a corresponding reduction in both circulating LDL-C and triglycerides as well as accumulation of neutral lipids in hepatocytes that can cause hepatic steatosis. Familial hypobetalipoproteinemia type 2 (also called familial combined hypolipidemia) is caused by loss of function mutations in both copies of the *ANGPTL3* gene. ANGPTL3 regulates lipolysis, so the loss of its activity results in increased extracellular lipolysis, increased clearance of all lipoproteins, and decreased concentrations of circulating lipids. The extent of cholesterol, triglyceride, and lipoprotein decrease is dependent on the severity of the mutations and the residual functionality of either ApoB or ANGPTL3.

Abetalipoproteinemia

Abetalipoproteinemia is a related condition caused by mutations in the microsomal triglyceride transfer protein, which is required for proper packaging and secretion of apoB-containing lipoproteins. Patients with the most severe mutations of these very low lipid disorders may present early in live with chronic diarrhea and failure to thrive, which is often not recognized as a lipid disorder despite having very low triglycerides, LDL-C, and apoB serum concentrations.

Hypoalphalipoproteinemia

HDL is protective against the damaging effects cholesterol deposition in arteries; individuals with low HDL are more susceptible to developing CVD. There are several etiologies that can result in hypoalphalipoproteinemia with extremely low concentrations (<20 mg/dL) of HDL-C (Table 28.2) [77]. Secondary causes of low HDL include insulin resistance, hypertriglyceridemia, obesity, smoking, liver disease, poor diet, malignancy, and certain prescription medications. The high concentration of paraproteins in patients with immunoproliferative disorders can cause a falsely low result on some homogeneous lipid assays via an interference with the lipoprotein blocking reagents used in those assays.

Familial hypoalphaproteinemia is rare (<1% of the general population) and results from one or more mutations in any of three main proteins required for either the nascent HDL formation or for reverse cholesterol transport. First, homozygous deletion mutations in both alleles for ApoA1 (the primary HDL lipoprotein) result in no reverse cholesterol efflux, very low HDL-C (<5 mg/dL), undetectable ApoA1, and planar xanthomas from cholesterol accumulation. Individuals heterozygous for the ApoA1 deletion have HDL-C and ApoA1 concentrations at approximately 50% of normal, whereas individuals with structural variants of ApoA1 can have variable concentrations of both HDL-C and ApoA1 (depending on the severity of the abnormality). Second, individuals with inactivating mutations in both genes encoding ABCA1 have a condition called Tangier disease (named after the island on which the first two cases were identified). Tangier disease results in low HDL-C and ApoA1 due to rapid clearance of nascent poorly lipidated HDL particles, and patients present with mild corneal opacification, hepatosplenomegaly, and enlarged orange tonsils. Individuals who are partially deficient for this protein have a milder presentation. Third, extremely low HDL-C is observed in individuals with complete LCATD, and the lipoprotein-associated enzyme requires for esterification of free cholesterol [78]. LCAT is synthesized mainly in the liver and is primarily associated with HDL particles, while a smaller proportion associates with apoB-containing particles. The enzyme esterifies free cholesterol at the surface of particles by transferring phosphatidylcholine of lecithin to cholesterol. Once esterified, the cholesterol migrates to the center of the particle and the particle increases in size. Consequently, patients with LCATD have an increased concentration of unesterified cholesterol (usually 60%−80% of plasma cholesterol is esterified in healthy individuals but it is <20% with LCATD) and very small nascent HDL particles that are rapidly cleared from circulation. Patients may also present with corneal opacification, chronic proteinuria, and renal failure. LCATD (or familial LCATD; FLD) is rare with a prevalence of <1:1,000,000. Partial LCATD, called fish eye disease, also results in very low HDL-C and corneal opacity but partial LCAT function is preserved.

Finally, chylomicron retention disease is another congenital disorder that results in little to no postprandial chylomicron formation, low total cholesterol, low HDL-C, and ApoA1 [79]. The causative gene is *SAR1B*, which encodes a GTPase that is required for prechylomicron trafficking from the endoplasmic reticulum to the Golgi within enterocytes of the intestine. The inability to produce chylomicrons results in fat-laden enterocytes, chronic steatorrhea, vomiting, abdominal distension, and delayed infant growth. The decreased HDL-C and ApoA1 is likely secondary from lipid droplet-induced intestinal damage given that approximately half of ApoA1 is synthesized in the intestine. Treatment for this disease is mainly limiting dietary fat intake while providing appropriate nutritional supplements.

TABLE 28.2 Potential causes of extremely low high-density lipoprotein cholesterol (<20 mg/dL).

Primary (monogenic) causes
ApoA1 deficiency
ApoA1 structural mutation
ABCA1 deficiency (Tangier disease)
LCAT deficiency
Secondary causes
Medications (anabolic androgenic steroids, fibrates, and thiazolidinedione in combination with fibrates)
Malignancy
Liver disease
Artifactual causes
Paraproteinemia interference with homogeneous direct lipid assay

ABCA1, Adenosine triphosphate binding cassette A1; *ApoA1*, apolipoprotein A-I; *HDL*, high-density lipoprotein; *LCAT*, lecithin: cholesterol acyltransferase.

Lipids and cardiovascular disease

Atherosclerosis is a disease in which plaque builds up within blood vessels. Atherosclerotic lesions begin when lipoproteins accumulate within arterial intima (Fig. 28.8) [80]. The rate of deposition within the vessel is directly proportional to the concentration of lipoproteins [81,82]. Cholesterol content of lipoproteins is a surrogate for lipoprotein concentration; however, the cholesterol content of

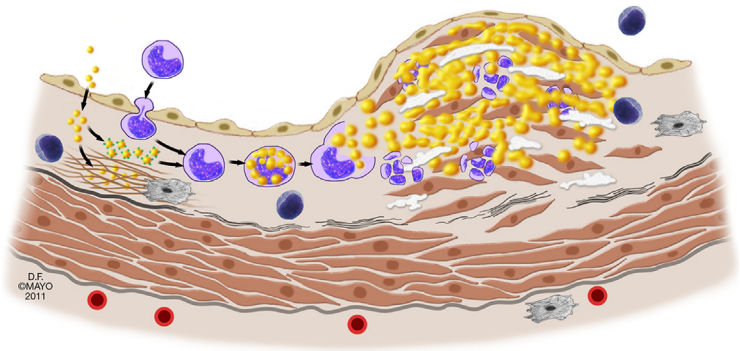


FIGURE 28.8 The role of lipids in atherosclerosis. Atherosclerosis begins when lipoproteins cross the vascular endothelium. Monocytes (turned macrophage) are drawn into the vascular intima to phagocytose the lipids. Macrophages overburdened with lipids become foam cells. Necrotic death of foam cells releases more cytokines, leading to inflammation and anatomical dysfunction of the atherosclerotic lesion.

lipoproteins is highly variable [83]. For these reasons, direct measures of lipoprotein concentrations such as APOB or LDL particle have been shown to give better performance in assessing a patient's risk of a future cardiovascular events compared with risk stratification based on LDL-C alone [33,84–92]. ApoB-lipoproteins are considered atherogenic lipoproteins [10,93]. Slow clearance of IDL and RLP leads to accumulation of LDL in the circulating plasma, which increases the probability of lipoproteins infiltrating the vascular endothelium and becoming trapped within the arterial intima [81,94].

Lipid lowering for treatment of cardiovascular disease

The assembly and secretion of VLDL by hepatocytes (or chylomicrons by enterocytes) is a complex multistep process. ApoB synthesis occurs at a relatively constant rate. Consequently, the rate of lipoprotein formation is a function of triglyceride and cholesterol availability. After implementing lifestyle modifications to lower lipid concentrations, several mechanisms can be pharmacologically targeted to reduce the concentration of circulating cholesterol and corresponding apoB lipoproteins, including inhibition of cholesterol synthesis in hepatocytes (statins), inhibition of cholesterol uptake in the gastro-intestinal system (Ezetimibe or bile acid sequestrants), or modulation of the apoB receptor concentration (PCSK9 inhibitors such as alirocumab or evolocumab).

The most prescribed agents are statins such as simvastatin, lovastatin, atorvastatin, or rosuvastatin. Statins inhibit 3-hydroxy 3-methylglutaryl CoA (HMG-CoA) reductase, the rate-limiting enzyme in human cholesterol synthesis [95]. Metaanalysis based on individual data from randomized clinical trials comparing statin medications to placebo found a mean reduction of 42 mg/dL LDL cholesterol among 175,000 individuals [96]. Modern treatment guidelines recommend statin doses, achieving a 30% or greater reduction in LDL cholesterol for at risk individuals [97]. According to National Health and Nutrition Examination Survey data, a third of U.S. adults ≥ 40 years of age are eligible for guideline-endorsed statin therapy [98,99]. Furthermore, the median LDL cholesterol among eligible adults is 105 mg/dL [100], suggesting that LDL cholesterol values <70 mg/dL will become increasingly common.

Aggressive LDL cholesterol lowering is unarguably beneficial to reducing cardiovascular events; however, it presents two challenges with respect to LDL cholesterol assessment by clinical laboratorians. The first challenge is the pervasive use of estimated LDL cholesterol by the Friedewald equation, which significantly underestimates actual LDL cholesterol at lower concentrations [101–103]. The second issue involves altered cholesterol composition of LDL following statin treatment. Multiple clinical trials have found that statin doses that reduce LDL cholesterol by 50% only reduce apoB by 36% [32,104–107]. These findings suggest that on-treatment subjects have cholesterol-depleted LDL. Since estimated and direct homogeneous LDL-C methods rely on certain assumptions about the relative amounts of triglycerides and cholesterol in lipoprotein fractions, it is reasonable to hypothesize that they would be less accurate in statin-treated subjects. Indeed, multiple studies have shown that on-treatment LDL-C is not significantly linked to the risk of CVD [108].

Clinical practice lipid measurement guidelines

Most laboratory measurements are interpreted in relation to a reference interval defined by the analyte concentration in a healthy reference population. However, interpretations of lipid results are most commonly guideline-driven from population data demonstrating risk of developing future CVD corresponding to various lipid concentrations. Values for desirable, borderline risk, and high risk were defined by the NCEP Adult Treatment Panel III [5], which was last updated in 2004. Additional pediatric and adolescent guidelines have been published for the interpretation of values and suggested screening intervals with an emphasis on early identification of children with FH (Table 28.3) [109].

TABLE 28.3 Basic lipid reference ranges endorsed by the National Cholesterol Education Program.

	Adults (≥18 year)		Pediatrics (<18 year)		
Lipid	**Value**	**Comment**	**Value**		**Comment**
LDL-C					
	<100	Desirable	<110		Acceptable
	100–129	Above desirable	110–129		Borderline high
	130–159	Borderline high	≥130		High
	160–189	High			
	>190	Very high			
Non-HDL-C					
	<130	Desirable	<120		Acceptable
	130–159	Above desirable	120–144		Borderline high
	160–189	Borderline high	≥145		High
	190–219	High			
	>220	Very high			
Total cholesterol					
	<200	Desirable	<170		Acceptable
	200–239	Borderline high	170–199		Borderline high
	>240	High	>200		High
HDL-C					
	<40 (men)	Low	<40		Low
	<50 (women)	Low	40–45		Borderline low
			>45		Acceptable
Triglycerides					
	<150	Normal	0–10 year	10–18 year	
	150–199	Borderline high	<75	<90	Acceptable
	200–499	High	75–99	90–129	Borderline high
	≥500	Very high	>100	>130	High

LDL-C, Low-density lipoprotein cholesterol; *HDL-C,* high-density lipoprotein cholesterol.

It is important to note that the NCEP no longer publishes guidelines and that clinical reference values for lipids are no longer considered therapeutic targets. Recommendations about risk assessment are now published by the AHA and ACC. Evidence from large randomized clinical trials has repeatedly shown that lowering cholesterol lowers risk of CVD, regardless of baseline cholesterol. These findings prompted recommendations that cholesterol lowering medications be prescribed to prevent future CVD based on patient risk factors. The highest dose of statins is recommended for individuals with a history of prior heart disease or LDL-C ≥ 190 mg/dL. Moderate doses are recommended for subjects with LDL-C between 70 and 190 mg/dL and diabetes mellitus or a 10-year risk >7.5% based on the pooled cohort calculation, which accounts for age, sex, race, blood pressure, lipids, and smoking history [6] (http://tools.acc.org/ASCVD-Risk-Estimator-Plus/).

References

[1] C. Cobbaert, et al., Reference standardization and triglyceride interference of a new homogeneous HDL-cholesterol assay compared with a former chemical precipitation assay, Clin. Chem. 44 (4) (1998) 779–789.

[2] N. Rifai, et al., Assessment of interlaboratory performance in external proficiency testing programs with a direct HDL-cholesterol assay, Clin. Chem. 44 (7) (1998) 1452–1458.

[3] W. Zhao, et al., Overestimation of HDL-cholesterol using a homogeneous "direct" assay, J. Clin. Lab. Anal. 18 (1) (2004) 42–44.

[4] Y. Ueda, et al., New homogeneous HDL-cholesterol assay without the influence of high TG sample using the selective detergent to lipoproteins, J. Clin. Lab. Anal. 17 (6) (2003) 201–208.

[5] Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, Executive summary of the third report of the National Cholesterol Education Program (NCEP) Expert Panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III), JAMA 285 (19) (2001) 2486–2497.

[6] S.M. Grundy, et al., 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines, J. Am. Coll. Cardiol. 73 (24) (2019) 3168–3209.

[7] J.H. Contois, G.R. Warnick, A.D. Sniderman, Reliability of low-density lipoprotein cholesterol, non-high-density lipoprotein cholesterol, and apolipoprotein B measurement, J. Clin. Lipidol. 5 (4) (2011) 264–272.

[8] P.S. Bachorik, J.W. Ross, National Cholesterol Education Program recommendations for measurement of low-density lipoprotein cholesterol: executive summary. The National Cholesterol Education Program Working Group on Lipoprotein Measurement, Clin. Chem. 41 (10) (1995) 1414–1420.

[9] R.J. Havel, E. Rapaport, Management of primary hyperlipidemia, N. Engl. J. Med. 332 (22) (1995) 1491–1498.

[10] B.G. Nordestgaard, A. Tybjaerg-Hansen, IDL, VLDL, chylomicrons and atherosclerosis, Eur. J. Epidemiol. 8 (Suppl. 1) (1992) 92–98.

[11] W.T. Friedewald, R.I. Levy, D.S. Fredrickson, Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge, Clin. Chem. 18 (6) (1972) 499–502.

[12] S.S. Martin, et al., Comparison of a novel method vs the Friedewald equation for estimating low-density lipoprotein cholesterol levels from the standard lipid profile, JAMA 310 (19) (2013) 2061–2068.

[13] P.H. Frost, R.J. Havel, Rationale for use of non-high-density lipoprotein cholesterol rather than low-density lipoprotein cholesterol as a tool for lipoprotein cholesterol screening and assessment of risk and therapy, Am. J. Cardiol. 81 (4A) (1998) 26B–31B.

[14] K.R. Jun, et al., Effects of total cholesterol and triglyceride on the percentage difference between the low-density lipoprotein cholesterol concentration measured directly and calculated using the Friedewald formula, Clin. Chem. Lab. Med. 46 (3) (2008) 371–375.

[15] J. Rubies-Prat, et al., Calculated low-density lipoprotein cholesterol should not be used for management of lipoprotein abnormalities in patients with diabetes mellitus, Diabetes Care 16 (8) (1993) 1081–1086.

[16] M. Senti, et al., Influence of intermediate-density lipoproteins on the accuracy of the Friedewald formula, Clin. Chem. 37 (8) (1991) 1394–1397.

[17] S.P. Whelton, et al., Evaluating the atherogenic burden of individuals with a Friedewald-estimated low-density lipoprotein cholesterol <70 mg/dL compared with a novel low-density lipoprotein estimation method, J. Clin. Lipidol. 11 (4) (2017) 1065–1072.

[18] J.W. Meeusen, et al., Reliability of calculated low-density lipoprotein cholesterol, Am. J. Cardiol. 116 (4) (2015) 538–540.

[19] J.W. Meeusen, et al., Validation of a proposed novel equation for estimating LDL cholesterol, Clin. Chem. 60 (12) (2014) 1519–1523.

[20] M. Sampson, C. Ling, Q. Sun, et al., A new equation for calculation of low-density lipoprotein cholesterol in patients with normolipidemia and/or hypertriglyceridemia, *JAMA Cardiol*. 2020. <10.1001/jamacardio.2020.0013>.

[21] N. Rifai, et al., Analytical and clinical performance of a homogeneous enzymatic LDL-cholesterol assay compared with the ultracentrifugation-dextran sulfate-Mg2 + method, Clin. Chem. 44 (6 Pt 1) (1998) 1242–1250.

[22] W.G. Miller, et al., Performance of four homogeneous direct methods for LDL-cholesterol, Clin. Chem. 48 (3) (2002) 489–498.

[23] H. Fei, et al., Evaluation of two different homogeneous assays for LDL-cholesterol in lipoprotein-X-positive serum, Clin. Chem. 46 (9) (2000) 1351–1356.

[24] S. Usui, et al., Differential reactivity of two homogeneous LDL-cholesterol methods to LDL and VLDL subfractions, as demonstrated by ultracentrifugation and HPLC, Clin. Chem. 48 (11) (2002) 1946–1954.

[25] P. Greenspan, et al., Advances in agarose gel electrophoresis of serum lipoproteins, J. Chromatogr. A 698 (1–2) (1995) 333–339.

[26] J.H. Contois, et al., Apolipoprotein B and cardiovascular disease risk: position statement from the AACC Lipoproteins and Vascular Diseases Division Working Group on Best Practices, Clin. Chem. 55 (3) (2009) 407–419.

[27] C.R. Harper, T.A. Jacobson, Using apolipoprotein B to manage dyslipidemic patients: time for a change? Mayo Clin. Proc. 85 (5) (2010) 440–445.

[28] J.D. Brunzell, et al., Lipoprotein management in patients with cardiometabolic risk: consensus conference report from the American Diabetes Association and the American College of Cardiology Foundation, J. Am. Coll. Cardiol. 51 (15) (2008) 1512–1524.

[29] S.M. Marcovina, et al., International Federation of Clinical Chemistry standardization project for measurements of apolipoproteins A-I and B, Clin. Chem. 37 (10) (1991) 1676–1682. Pt 1.

[30] S.M. Marcovina, et al., International Federation of Clinical Chemistry standardization project for measurements of apolipoproteins A-I and B. IV. Comparability of apolipoprotein B values by use of International Reference Material, Clin. Chem. 40 (4) (1994) 586–592.

[31] European Association for Cardiovascular Prevention & Rehabilitation, et al., ESC/EAS Guidelines for the management of dyslipidaemias: the Task Force for the management of dyslipidaemias of the European Society of Cardiology (ESC) and the European Atherosclerosis Society (EAS), Eur. Heart J. 32 (14) (2011) 1769–1818.

[32] M.J. McQueen, et al., Lipids, lipoproteins, and apolipoproteins as risk markers of myocardial infarction in 52 countries (the INTERHEART study): a case-control study, Lancet 372 (9634) (2008) 224−233.
[33] A. Thompson, J. Danesh, Associations between apolipoprotein B, apolipoprotein AI, the apolipoprotein B/AI ratio and coronary heart disease: a literature-based meta-analysis of prospective studies, J. Intern. Med. 259 (5) (2006) 481−492.
[34] B.G. Nordestgaard, et al., Fasting is not routinely required for determination of a lipid profile: clinical and laboratory implications including flagging at desirable concentration cutpoints-a joint consensus statement from the European Atherosclerosis Society and European Federation of Clinical Chemistry and Laboratory Medicine, Clin. Chem. 62 (7) (2016) 930−946.
[35] M. Scartezini, et al., Positioning about the flexibility of fasting for lipid profiling, Arq. Bras. Cardiol. 108 (3) (2017) 195−197.
[36] P.S. Jellinger, et al., American Association of Clinical Endocrinologists and American College of Endocrinology Guidelines for Management of Dyslipidemia and Prevention of Cardiovascular Disease, Endocr. Pract. 23 (Suppl. 2) (2017) 1−87.
[37] A.L. Catapano, et al., 2016 ESC/EAS guidelines for the management of dyslipidaemias, Eur. Heart J. 37 (39) (2016) 2999−3058.
[38] T.J. Anderson, et al., 2016 Canadian Cardiovascular Society Guidelines for the management of dyslipidemia for the prevention of cardiovascular disease in the adult, Can. J. Cardiol. 32 (11) (2016) 1263−1282.
[39] J.R. Downs, P.G. O'Malley, Management of dyslipidemia for cardiovascular disease risk reduction: synopsis of the 2014 U.S. Department of Veterans Affairs and U.S. Department of Defense clinical practice guideline, Ann. Intern. Med. 163 (4) (2015) 291−297.
[40] G.M. Egeland, et al., Non-fasting triglycerides predict incident acute myocardial infarction among those with favourable HDL-cholesterol: Cohort Norway, Eur. J. Prev. Cardiol. 22 (7) (2015) 872−881.
[41] A. Langsted, J.J. Freiberg, B.G. Nordestgaard, Fasting and nonfasting lipid levels: influence of normal food intake on lipids, lipoproteins, apolipoproteins, and cardiovascular risk prediction, Circulation 118 (20) (2008) 2047−2056.
[42] V. Sathiyakumar, et al., Fasting vs non-fasting and low-density lipoprotein-cholesterol accuracy, Circulation (2017).
[43] B.G. Nordestgaard, A test in context: lipid profile, fasting versus nonfasting, J. Am. Coll. Cardiol. 70 (13) (2017) 1637−1646.
[44] S.M. Marcovina, V.P. Gaur, J.J. Albers, Biological variability of cholesterol, triglyceride, low- and high-density lipoprotein cholesterol, lipoprotein(a), and apolipoproteins A-I and B, Clin. Chem. 40 (4) (1994) 574−578.
[45] L.J. Cartier, et al., Comparison of fasting and non-fasting lipid profiles in a large cohort of patients presenting at a community hospital, Clin. Biochem. 52 (2018) 61−66.
[46] J.W. Meeusen, Comparing measures of HDL: on the right path with the wrong map, Clin. Chem. 64 (3) (2018) 424−426.
[47] R.C. Hoogeveen, et al., Small dense low-density lipoprotein-cholesterol concentrations predict risk for coronary heart disease: the Atherosclerosis Risk In Communities (ARIC) study, Arterioscler. Thromb. Vasc. Biol. 34 (5) (2014) 1069−1077.
[48] F.M. Sacks, H. Campos, Clinical review 163: cardiovascular endocrinology: low-density lipoprotein size and cardiovascular disease: a reappraisal, J. Clin. Endocrinol. Metab. 88 (10) (2003) 4525−4532.
[49] S. Ip, et al., Systematic review: association of low-density lipoprotein subfractions with cardiovascular outcomes, Ann. Intern. Med. 150 (7) (2009) 474−484.
[50] S. Mora, Advanced lipoprotein testing and subfractionation are not (yet) ready for routine clinical use, Circulation 119 (17) (2009) 2396−2404.
[51] Biochemistry, N.A.O.C., Emerging Biomarkers for Primary Prevention of Cardiovascular Disease and Stroke, AACC, 2009. <https://www.aacc.org/science-and-practice/practice-guidelines/emerging-cv-risk-factors>.
[52] M.H. Davidson, et al., Clinical utility of inflammatory markers and advanced lipoprotein testing: advice from an expert panel of lipid specialists, J. Clin. Lipidol. 5 (5) (2011) 338−367.
[53] P. Greenland, et al., 2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, Circulation 122 (25) (2010) e584−e636.
[54] D.S. Fredrickson, R.S. Lees, A system for phenotyping hyperlipoproteinemia, Circulation 31 (1965) 321−327.
[55] A. Brahm, R.A. Hegele, Hypertriglyceridemia, Nutrients 5 (3) (2013) 981−1001.
[56] M. Miller, et al., Triglycerides and cardiovascular disease: a scientific statement from the American Heart Association, Circulation 123 (20) (2011) 2292−2333.
[57] D.J. Blom, et al., Characterizing familial chylomicronemia syndrome: baseline data of the APPROACH study, J. Clin. Lipidol. 12 (5) (2018) 1234−1243 e5.
[58] L. Berglund, et al., Evaluation and treatment of hypertriglyceridemia: an Endocrine Society clinical practice guideline, J. Clin. Endocrinol. Metab. 97 (9) (2012) 2969−2989.
[59] H. Nawaz, et al., Elevated serum triglycerides are independently associated with persistent organ failure in acute pancreatitis, Am. J. Gastroenterol. 110 (10) (2015) 1497−1503.
[60] L.E. Akioyamen, et al., Estimating the prevalence of heterozygous familial hypercholesterolaemia: a systematic review and meta-analysis, BMJ Open. 7 (9) (2017) e016461.
[61] E. Leitersdorf, et al., Common low-density lipoprotein receptor mutations in the French Canadian population, J. Clin. Invest. 85 (4) (1990) 1014−1023.
[62] B. Sjouke, et al., Homozygous autosomal dominant hypercholesterolaemia: prevalence, diagnosis, and current and future treatment perspectives, Curr. Opin. Lipidol. 26 (3) (2015) 200−209.
[63] M.M. van Greevenbroek, et al., Familial combined hyperlipidemia: from molecular insights to tailored therapy, Curr. Opin. Lipidol. 25 (3) (2014) 176−182.
[64] P. Mata, et al., [Familial combined hyperlipidemia: consensus document], Semergen 40 (7) (2014) 374−380.
[65] P.N. Hopkins, E.A. Brinton, M.N. Nanjee, Hyperlipoproteinemia type 3: the forgotten phenotype, Curr. Atheroscler. Rep. 16 (9) (2014) 440.
[66] J. Morganroth, R.I. Levy, D.S. Fredrickson, The biochemical, clinical, and genetic features of type III hyperlipoproteinemia, Ann. Intern. Med. 82 (2) (1975) 158−174.
[67] M. Fainaru, et al., Structural and metabolic heterogeneity of beta-very low density lipoproteins from cholesterol-fed dogs and from humans with type III hyperlipoproteinemia, J. Lipid Res. 23 (5) (1982) 702−714.
[68] R.A. Hegele, et al., The polygenic nature of hypertriglyceridaemia: implications for definition, diagnosis, and management, Lancet Diabetes Endocrinol. 2 (8) (2014) 655−666.

[69] D.P. Wilson, et al., Use of lipoprotein(a) in clinical practice: a biomarker whose time has come. A scientific statement from the National Lipid Association, J. Clin. Lipidol. 13 (3) (2019) 374–392.
[70] S. Tsimikas, A test in context: lipoprotein(a): diagnosis, prognosis, controversies, and emerging therapies, J. Am. Coll. Cardiol. 69 (6) (2017) 692–711.
[71] Emerging Risk Factors Collaboration, et al., Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality, JAMA 302 (4) (2009) 412–423.
[72] B.G. Nordestgaard, et al., Lipoprotein(a) as a cardiovascular risk factor: current status, Eur. Heart J. 31 (23) (2010) 2844–2853.
[73] S.M. Marcovina, et al., Use of a reference material proposed by the International Federation of Clinical Chemistry and Laboratory Medicine to evaluate analytical methods for the determination of plasma lipoprotein(a), Clin. Chem. 46 (12) (2000) 1956–1967.
[74] A.V. Khera, et al., Lipoprotein(a) concentrations, rosuvastatin therapy, and residual vascular risk: an analysis from the JUPITER Trial (Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin), Circulation 129 (6) (2014) 635–642.
[75] L.A. Freeman, et al., Plasma lipoprotein-X quantification on filipin-stained gels: monitoring recombinant LCAT treatment ex-vivo, J. Lipid Res. (2019).
[76] A. Di Costanzo, et al., Clinical and biochemical characteristics of individuals with low cholesterol syndromes: a comparison between familial hypobetalipoproteinemia and familial combined hypolipidemia, J. Clin. Lipidol. 11 (5) (2017) 1234–1242.
[77] D.J. Rader, E.M. deGoma, Approach to the patient with extremely low HDL-cholesterol, J. Clin. Endocrinol. Metab. 97 (10) (2012) 3399–3407.
[78] R. Saeedi, M. Li, J. Frohlich, A review on lecithin:cholesterol acyltransferase deficiency, Clin. Biochem. 48 (7–8) (2015) 472–475.
[79] E. Levy, P. Poinsot, S. Spahis, Chylomicron retention disease: genetics, biochemistry, and clinical spectrum, Curr. Opin. Lipidol. 30 (2) (2019) 134–139.
[80] I. Tabas, K.J. Williams, J. Boren, Subendothelial lipoprotein retention as the initiating process in atherosclerosis: update and therapeutic implications, Circulation 116 (16) (2007) 1832–1844.
[81] M. Shaikh, et al., Quantitative studies of transfer in vivo of low density, Sf 12–60, and Sf 60–400 lipoproteins between plasma and arterial intima in humans, Arterioscler. Thromb. Vasc. Biol. 11 (3) (1991) 569–577.
[82] S. Stender, D.B. Zilversmit, Transfer of plasma lipoprotein components and of plasma proteins into aortas of cholesterol-fed rabbits. Molecular size as a determinant of plasma lipoprotein influx, Arteriosclerosis 1 (1) (1981) 38–49.
[83] B. Teng, et al., Composition and distribution of low density lipoprotein fractions in hyperapobetalipoproteinemia, normolipidemia, and familial hypercholesterolemia, Proc. Natl Acad. Sci. U. S. A. 80 (21) (1983) 6662–6666.
[84] G. Walldius, et al., High apolipoprotein B, low apolipoprotein A-I, and improvement in the prediction of fatal myocardial infarction (AMORIS study): a prospective study, Lancet 358 (9298) (2001) 2026–2033.
[85] A.C. St-Pierre, et al., Low-density lipoprotein subfractions and the long-term risk of ischemic heart disease in men: 13-year follow-up data from the Quebec Cardiovascular Study, Arterioscler. Thromb. Vasc. Biol. 25 (3) (2005) 553–559.
[86] P.J. Talmud, et al., Nonfasting apolipoprotein B and triglyceride levels as a useful predictor of coronary heart disease risk in middle-aged UK men, Arterioscler. Thromb. Vasc. Biol. 22 (11) (2002) 1918–1923.
[87] I. Shai, et al., Multivariate assessment of lipid parameters as predictors of coronary heart disease among postmenopausal women: potential implications for clinical guidelines, Circulation 110 (18) (2004) 2824–2830.
[88] A.R. Sharrett, et al., Coronary heart disease prediction from lipoprotein cholesterol levels, triglycerides, lipoprotein(a), apolipoproteins A-I and B, and HDL density subfractions: The Atherosclerosis Risk in Communities (ARIC) Study, Circulation 104 (10) (2001) 1108–1113.
[89] A.M. Gotto Jr., et al., Relation between baseline and on-treatment lipid parameters and first acute major coronary events in the Air Force/Texas Coronary Atherosclerosis Prevention Study (AFCAPS/TexCAPS), Circulation 101 (5) (2000) 477–484.
[90] P.M. Ridker, et al., Non-HDL cholesterol, apolipoproteins A-I and B100, standard lipid measures, lipid ratios, and CRP as risk factors for cardiovascular disease in women, JAMA 294 (3) (2005) 326–333.
[91] M.J. Stampfer, et al., A prospective study of cholesterol, apolipoproteins, and the risk of myocardial infarction, N. Engl. J. Med. 325 (6) (1991) 373–381.
[92] M. Benn, et al., Improving prediction of ischemic cardiovascular disease in the general population using apolipoprotein B: the Copenhagen City Heart Study, Arterioscler. Thromb. Vasc. Biol. 27 (3) (2007) 661–670.
[93] A. Varbo, J.J. Freiberg, B.G. Nordestgaard, Extreme nonfasting remnant cholesterol vs extreme LDL cholesterol as contributors to cardiovascular disease and all-cause mortality in 90000 individuals from the general population, Clin. Chem. 61 (3) (2015) 533–543.
[94] B.G. Nordestgaard, L.B. Nielsen, Atherosclerosis and arterial influx of lipoproteins, Curr. Opin. Lipidol 5 (4) (1994) 252–257.
[95] S.M. Grundy, HMG-CoA reductase inhibitors for treatment of hypercholesterolemia, N. Engl. J. Med. 319 (1) (1988) 24–33.
[96] B. Mihaylova, et al., The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials, Lancet 380 (9841) (2012) 581–590.
[97] N.J. Stone, et al., 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines, Circulation 129 (25 Suppl. 2) (2014) S1–S45.
[98] D. Mozaffarian, et al., Heart disease and stroke statistics--2015 update: a report from the American Heart Association, Circulation 131 (4) (2015) e29–e322.
[99] D.C. Goff Jr., et al., 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines, Circulation 129 (25 Suppl. 2) (2014) S49–S73.
[100] M.J. Pencina, et al., Application of new cholesterol guidelines to a population-based sample, N. Engl. J. Med. 370 (15) (2014) 1422–1431.

[101] H. Scharnagl, et al., The Friedewald formula underestimates LDL cholesterol at low concentrations, Clin. Chem. Lab. Med. 39 (5) (2001) 426–431.
[102] L. Sibal, et al., Friedewald equation underestimates low-density lipoprotein cholesterol at low concentrations in young people with and without Type 1 diabetes, Diabet. Med. 27 (1) (2010) 37–45.
[103] S.S. Martin, et al., Friedewald-estimated versus directly measured low-density lipoprotein cholesterol and treatment implications, J. Am. Coll. Cardiol. 62 (8) (2013) 732–739.
[104] S.M. Boekholdt, et al., Very low levels of atherogenic lipoproteins and the risk for cardiovascular events: a meta-analysis of statin trials, J. Am. Coll. Cardiol. 64 (5) (2014) 485–494.
[105] C.M. Ballantyne, J.S. Raichlen, V.A. Cain, Statin therapy alters the relationship between apolipoprotein B and low-density lipoprotein cholesterol and non-high-density lipoprotein cholesterol targets in high-risk patients: the MERCURY II (Measuring Effective Reductions in Cholesterol Using Rosuvastatin) trial, J. Am. Coll. Cardiol. 52 (8) (2008) 626–632.
[106] P.H. Jones, et al., Effects of rosuvastatin versus atorvastatin, simvastatin, and pravastatin on non-high-density lipoprotein cholesterol, apolipoproteins, and lipid ratios in patients with hypercholesterolemia: additional results from the STELLAR trial, Clin. Ther. 26 (9) (2004) 1388–1399.
[107] N.D. Wong, et al., Residual dyslipidemia according to low-density lipoprotein cholesterol, non-high-density lipoprotein cholesterol, and apolipoprotein B among statin-treated US adults: National Health and Nutrition Examination Survey 2009–2010, J. Clin. Lipidol. 9 (4) (2015) 525–532.
[108] C.M. Ballantyne, et al., Apolipoprotein E genotypes and response of plasma lipids and progression-regression of coronary atherosclerosis to lipid-lowering drug therapy, J. Am. Coll. Cardiol. 36 (5) (2000) 1572–1578.
[109] Expert Panel on Integrated Guidelines for Cardiovascular Health and Risk Reduction in Children and Adolescents, et al., Expert panel on integrated guidelines for cardiovascular health and risk reduction in children and adolescents: summary report, Pediatrics 128 (Suppl. 5) (2011) S213–S256.

Self-assessment questions

1. Correctly list major lipoproteins in order from largest to smallest:
   a. VLDL, chylomicrons, IDL, HDL, LDL
   b. Chylomicrons, VLDL, IDL, HDL, LDL
   c. VLDL, IDL, chylomicrons, LDL, HDL
   d. Chylomicrons, VLDL, IDL, LDL, HDL
2. What is the calculated non-HDL cholesterol for a patient with the following lipid values: total cholesterol = 226 mg/dL, HDL cholesterol = 37 mg/dL, and triglycerides = 113 mg/dL?
   a. 76 mg/dL
   b. 189 mg/dL
   c. 166 mg/dL
   d. 150 mg/dL
3. Which laboratory measure below is most affected by the patient's fasting status?
   a. Total Cholesterol
   b. HDL Cholesterol
   c. Triglycerides
   d. Apolipoprotein B
4. Explain the rationale behind the requirement of fasting for lipid assessment.
   a. HDL cholesterol decreases changes after a meal
   b. LDL cholesterol decreases after a meal
   c. VLDL triglycerides increase after a meal
   d. Chylomicron triglycerides increase after a meal
5. Statin drugs (lovastatin, simvastatin, pravastatin, etc.) lower cholesterol by:
   a. activating HMG-CoA reductase
   b. activating lipoprotein lipase
   c. inhibiting HMG-CoA reductase
   d. inhibiting lipoprotein lipase
6. You are asked to consult on a case involving a 4-year-old boy who presents with xanthomas on his arms and legs. Laboratory results reveal total cholesterol of 740 mg/dL, an LDL-C of 240 mg/dL, an HDL-C of 52 mg/dL, and a triglyceride of 100 mg/dL. At first glance, you are suspicious of:
   a. FH
   b. lipoprotein lipase deficiency
   c. LCAT deficiency
   d. familial combined hyperlipidemia
7. For a patient who has an elevated Lp(a), the abnormality most likely to be detected by a basic lipid panel is:
   a. elevated VLDL cholesterol
   b. low HDL cholesterol
   c. elevated LDL cholesterol
   d. elevated fasting triglycerides
8. Which analyte is most appropriate to use as a surrogate for the number of LDL particles?
   a. Apolipoprotein A1
   b. Apolipoprotein B
   c. Apolipoprotein CIII
   d. Apolipoprotein E

Answers

1. d
2. b
3. c
4. d
5. c
6. a
7. c
8. b

Chapter 29

Pediatric laboratory medicine

Amy L. Pyle-Eilola[1] and Jane Dickerson[2]

[1]*Department of Pathology and Laboratory Medicine, Nationwide Children's Hospital, Columbus, OH, United States,* [2]*Department of Laboratory Medicine, University of Washington Medical Center, Seattle, WA, United States*

Learning objectives

After reviewing this chapter, the reader will be able to:

- Identify three challenges associated with pediatric sample volumes.
- Name the specimen types unique to pediatrics.
- Discuss the role of the clinical laboratory in newborn screening.
- Define the disease states typified by elevated bilirubin concentrations in early childhood.

"Children are not little adults" is a common phrase in the practice of pediatrics, and laboratory medicine is not exempt from issues that pertain primarily to the pediatric population. Handling pediatric specimens presents unique challenges ranging from small blood volumes, sample types, and modes of samples collection, to test validation and establishing reference intervals. Many of these issues affect the day-to-day processes in the clinical laboratory. One of the biggest problems in pediatrics is small sample volume, which affects almost every step in the process from instrument selection all the way to reference intervals and result interpretation.

Blood volume and sample volume requirements

The volume of blood that can be drawn safely per day from an individual is based on the individual's total body blood volume. The small total body blood volume of infants and children limits the amount of blood that can be drawn in this patient population. Current guidelines recommend that no more than 3% of total body volume be drawn per day from infants less than 2 months of age, and no more than 10% of total body volume be drawn per day in healthy children over 2 months of age (Fig. 29.1) [1]. Thus, while 150 mL of blood can be collected from a healthy adult in a day, only 9 mL of blood can be obtained safely from a healthy 7-lb infant. In the pediatric intensive care unit, blood draws for lab testing account for 73% of blood loss, and a single patient typically experiences 7.1 ± 5.3 mL of blood loss per day in the unit [2]. This daily loss closely approaches the recommended total loss for an entire month in a 7-lb infant. Often, to reduce the volume of blood drawn for laboratory testing, the minimum amount of blood necessary for testing is collected in smaller collection devices, rather than filling standard size collection devices.

The minimum volume of blood needed for testing is the sum of volume needed for the actual assay, the volume needed for sample integrity indices, and the instrument dead volume. The instrument dead volume is essentially wasted volume that is required for instrument function and is the volume below which the instrument cannot pipette. The dead volume of the instrument can be quite considerable when compared with the actual assay volume, which is typically very small. For example, an assay may require 30 μL, but the dead volume of the instrument may be 240 μL, bringing the total sample volume to 270 μL. Therefore it is important to be mindful when assessing a manufacturer's claims for sample volume requirements. The dead volume should be validated within the laboratory, as requirements may vary from the manufacturer's specifications. When validating the dead volume, the tube size that will be used most often should be tested. The dead volume may vary considerably between a 10-mL primary tube and a 0.5-mL sample cup. For assays requiring serum and plasma, the volume of whole blood has to be sufficient to provide the necessary volume of serum or plasma.

Hematocrit, or the fraction of red blood cells in the blood, will affect the amount of serum in a whole blood specimen. For example, in a specimen with a high hematocrit, a larger volume of whole blood is required to obtain the desired volume of serum. Neonates and infants have higher hematocrit, with the upper end of the reference interval close to 70%, whereas hematocrit in older

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00029-6

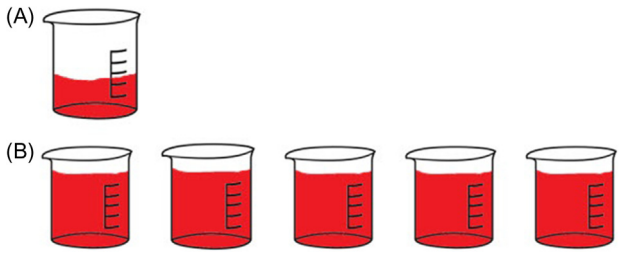


FIGURE 29.1 Illustration of the total blood volume of (A) an 8-lb baby (340 mL) and the total blood volume of (B) a 150-lb adult (~5 L).

children and adults is typically less than 50%. Whole blood assays require less blood overall, since plasma is roughly 40%–60% of a whole blood specimen. This is one reason that assays that require whole blood specimens are preferred for pediatric laboratory testing. Point-of-care systems commonly utilize very small volumes of whole blood and have little to no dead volume requirements, making them attractive options in pediatrics. For example, a whole blood glucose measurement can be obtained from a drop of blood (~10 μL), compared with 80–100 μL for plasma glucose testing in the central lab. However, there are disadvantages to consider with the use of whole blood point-of-care systems; see Chapter 19, Point-of-care testing, for more details.

Challenges associated with small blood volumes

Small blood volumes are prone to particular preanalytical errors. Common errors associated with small volume samples include incorrect blood-to-additive ratios, sample evaporation, hemolysis, and human error during manual specimen processing. Underfilled tubes lead to decreased blood-to-additive ratios, and incorrect ratios can give rise to inaccurate test results, hemolysis, and altered cell morphology. For example, coagulation assays require a specific concentration of anticoagulants, and in cases where sodium citrate tubes are underfilled, the abnormally high concentration of citrate will falsely prolong clotting times. Higher concentrations of additives contribute to other preanalytical problems such as hemolysis and changes to cell morphology. Red blood cells exposed to high concentrations of glycolytic inhibitors, like sodium fluoride, are more likely to hemolyze, and cells exposed to hyperosmolar concentrations of EDTA may shrink and exhibit an artifactual decreased MCV [3].

Evaporation is another major concern for small volume samples given that smaller volumes have a greater surface to total volume ratio. Evaporation of small volume samples can occur during or after sample processing, thus creating large differences in analyte concentrations. The effects of evaporation must be considered when adding on a test to an existing specimen. For example, a sample containing 2.0 mL of separated serum (a typical adult volume) that sits open to the air may show an increase in analyte concentration that is up to 10% higher than the initial measurement after 4 h. By contrast, a sample containing only 0.5 mL under the same conditions may show a 50% increase over the initial analysis. Even frozen samples are prone to evaporation when stored for long periods. To reduce the effects of evaporation, samples should be tightly capped any time they are not actively tested [4,5]. Volatile compounds like ethanol and dissolved gases, such as CO_2, are also significantly impacted by high surface to volume ratios. Total CO_2 (TCO_2) is typically measured in plasma via enzymatic techniques following alkali treatment. Alkali treatment quantitatively converts circulating forms of CO_2 (HCO_3^-, H_2CO_3, and dissolved CO_2) to HCO_3^-, which serves as the limiting substrate for phosphoenolpyruvate carboxykinase. At high surface to volume ratios, extended exposure of a small volume sample (pCO_2 ~40–50 mmHg) to air (pCO_2 ~0.3 mmHg) can rapidly deplete the pool of dissolved CO_2, reducing apparent TCO_2 concentrations. These changes are often incorrectly assumed to reflect metabolic acidosis in the patient, leading to unnecessary repeat testing and further metabolic testing to explain the suspected acidosis.

Small volume samples can be collected in reduced vacuum tubes or in microcollection tubes often referred to as "bullets." These tubes contain lower quantities of additives and are designed for collection of 0.5–1 mL of specimen. The use of these miniature collection tubes poses several obstacles to laboratory workflow, especially for highly automated laboratories. Bullets do not fit most analyzers and automated robotic systems. Standard labels are often too large for bullets and specialized labels, and printers may be required. Laboratories not equipped with such special printers have no choice but to manually enter patient information into the laboratory information system and analyzer. The addition of these manual processes to the workflow can introduce errors, prolong turnaround time, and influence the number of employees needed to manage such specimen types. One potential solution to this problem is to place the bullet into a larger tube that can be barcoded and positioned into the instrument. Some manufacturers make false-bottom tubes that fit on some instrument platforms, but more often, transfer of the sample to a compatible sample cup and relabeling are required prior to analysis. These additional steps introduce more opportunities for error. Measures taken to accommodate small volumes, such as false-bottom tubes, sample cups, or tube within a tube, all require validation prior to implementation into laboratory workflows.

Specimen types and collection

Small volume blood samples are often obtained via capillary blood collection from a finger or heel. Heel-stick collections are commonly performed in infants who are not walking. There are strict guidelines for obtaining heel-stick collections [6]. The heel must be warmed prior to sample collection, as this increases the blood flow to the area and changes the blood composition in the heel to better reflect the composition of arterial blood. The correct lancet length must be utilized. The longer tips result in deeper punctures and more freely flowing blood, but at the risk of puncturing the bone, which can lead to osteomyelitis. Devices with tips no longer than 2.0 mm are recommended for infant heel sticks, and most devices for use in newborns have tip lengths less than 1.0 mm. The heel must be pricked on the lateral or medial plantar surfaces of the heel, and not the back of the heel (Fig. 29.2). The bone is too close to the surface at the back of the heel, and hitting the bone would not only cause pain but also risk for osteomyelitis infections [7]. "Milking" (excessive and repetitive squeezing) the heel must be avoided to reduce sample contamination by tissue factor and to reduce the risk of sample hemolysis. Since exposure of slow flowing capillary blood to disrupted tissue can trigger clot formation, coagulation testing is contraindicated on capillary specimens. Furthermore, dependence on capillary samples may lead to inaccuracy and exaggerated variability. Cholesterol measurement in capillary blood was demonstrated to be positively biased compared with venous blood [8]. A recent study comparing WBC counts, three-part WBC differential, and platelet counts from successive drops of capillary blood reported the average drop-to-drop coefficient of variation was five times higher than obtained with well-mixed venous blood [9].

Urine can be challenging to collect in the pediatric population, especially in infants. Urine should never be collected by squeezing cotton balls, gauze, or diapers into the collection container. If not collected by catheterization, urine can be collected by urine collection bags that have adhesive surfaces and can be placed around the genitals; 24-h collections are best obtained via catheterization.

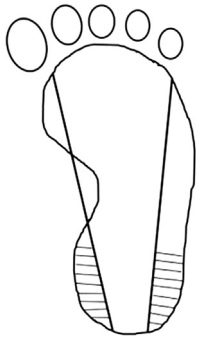


FIGURE 29.2 Capillary blood is safely collected from puncture of the lateral and medial sides of the heel (shaded portions), not the middle of the heel.

Specimen types unique to pediatrics include umbilical cord and meconium. Meconium is the first stool of a newborn infant. It starts to form about the $12^{th}-16^{th}$ week of gestation, and can be used to detect maternal drug use in the last 2–3 months of gestation, depending on the specific drug class. Because it accumulates during gestation, it is more sensitive for investigating long-term maternal drug use than urine, which will only be positive if the mother used drugs within days or weeks of delivery. Meconium can be collected by scraping the specimen from the infant's diaper. Disadvantages to meconium are that it may be passed in utero and therefore not be available to test, and that, when it is available after birth, it is not a homogenous sample and the specimen must be thoroughly mixed, so that all components are equally distributed prior to analysis. The umbilical cord has become more commonly used as an alternative specimen type to meconium, especially in instances where it is not possible to collect meconium, such as after meconium ileus [10]. While the analytical performance is comparable between meconium and umbilical cords, it is important to consider that the distribution of drugs and metabolites in either specimen type is not well understood and discordance has been observed between the two sample types. Most hospitals are not equipped to analyze meconium or umbilical cord tissue; thus these specimens are sent to alternative testing facilities. These specialty reference labs are not standardized in either their test menus (drugs/metabolites they detect) or assay performance. This provides an additional layer of difficulty in interpreting drug testing results in neonates.

Reference intervals in pediatrics

Reference intervals are established by using either parametric or nonparametric analyses and according to guidelines set forth by the Clinical Laboratory Standards Institute. Establishment of new reference intervals requires 120 healthy reference individuals for each partition (e.g., sex and age groups). In practice, it is more common that a lab verifies a previously established reference interval with their test results. Application of adult reference intervals is not always valid in pediatrics; this is true for all areas of laboratory testing including chemistry, hematology, immunology, and coagulation. Establishing appropriate reference ranges remains one of the foremost challenges in pediatric laboratory medicine.

Many available reference intervals for pediatrics were developed decades ago on outdated instrumentation, and many used hospitalized infants and children as reference

individuals [11]. Using healthy reference individuals is important when establishing intervals; however, recruiting healthy individuals less than 18 years of age and especially less than 5 years of age is difficult. The number needed to establish an interval is increased by the need to partition due to dynamic changes that occur during childhood, particularly during infancy and puberty. There are some large-scale initiatives to establish clinical laboratory reference intervals in pediatrics in both Canada and the United States: the Canadian Laboratory Initiative on Pediatric Reference Intervals database and the Children's Health Improvement through Laboratory Diagnostics initiative out of ARUP Laboratories, Salt Lake City, UT, United States [11,12]. In addition, the American Association for Clinical Chemistry has also worked with the US government on several initiatives and is currently advocating for a new nationwide reference interval program [13]. Interpreting results and use of reference intervals also requires attention. Dynamic reference intervals might confound results. For example, immunoreactive trypsinogen (IRT) rises quickly after birth and may lead to false positives on newborn screens if the screening sample is collected several days after birth. In addition, the result must be reviewed in clinical context. A patient who is followed over time may suddenly have normal or abnormal results after a birthday moves them into a new interval partition.

Newborn screening

Every US state and most industrialized countries have regulations in place to test newborns for various metabolic and nonmetabolic conditions, meeting defined criteria established by the World Health Organization. Conceptually, the criteria can be simplified to two main considerations: (1) newborn should benefit from early detection; and (2) the benefit should be balanced by the costs, both financial and risk of harm. With the discovery that mental retardation seen in untreated phenylketonuria (PKU) can be prevented by early dietary restriction of phenylalanine, and that affected patients can be detected by the elevation of phenylalanine in their blood, it became imperative to identify babies with PKU before the onset of developmental delays. This led to the introduction of newborn screening for PKU in 1961 and, subsequently, other conditions were included. Initially, amino acid disorders were tested using bacterial inhibition assays in which bacteria only grew when the concentration of a specific amino acid was elevated, called the Guthrie test. This limited testing to those conditions in which auxotrophic bacteria could identify an accumulating metabolite.

Blood for newborn screening samples is currently collected on a filter paper during the first week of life, ideally between 24−72 h. There are standard guidelines for the collection of newborn screening samples [14]. The filter paper is manufactured to absorb a standard amount of blood. This standardization allows approximation of the concentration of the analyte in plasma based on the volume of blood in the sample of paper used in the test. These approximations are reasonably accurate for term infants without anemia, but the concentrations are less accurate in premature infants with lower hematocrits and infants with polycythemia.

In addition to hematocrit, several other factors can affect the accuracy and specificity of the newborn screening results. These include gestational age, postnatal age, type of infant feeding, and the time of sample collection in relation to feeding. For this reason, the positive predictive value of an abnormal newborn screening may be as high as 1 in 3 or as low as 1 in 500; thus more definitive testing must follow abnormal screening results. Many state programs employ second-tier tests to augment the primary screening result, which improves the specificity and positive predictive value for those conditions. Recently, there has been a push toward redefining analyte cutoff values to be reflective of a disease state, rather than calculated from the central 95% of a "normal" population [15].

Inborn errors of metabolism

With the introduction of tandem mass spectrometry (MS/MS) and other technologies into the clinical diagnostic laboratory, newborn screening has now expanded to include amino acid disorders, organic acidemias, fatty acid and carnitine disorders, congenital adrenal hypoplasia, galactosemia, biotinidase deficiency, as well as hypothyroidism, severe combined immunodeficiency (SCID) disease, and other nonmetabolic conditions. Currently, most screening includes quantitation of selected amino acids and carnitine esters by MS/MS on the same sample.

Organic acidemias can also be detected by MS/MS via accumulation of specific acylcarnitine species (e.g., elevated C3 acylcarnitine may indicate methylmalonic academia or proprionic, and elevated C5 acylcarnitine may be the result of isovaleric acidemia). By using these methods, benign maternal organic acid disorders have been incidentally discovered at a higher rate than expected, including 3-methylcrotonyl CoA carboxylase deficiency and vitamin B12 deficiency [16].

Enzymatic measurement of biotinidase is performed as a semiquantitative test using an artificial substrate [biotinyl-*p*-aminobenzoic acid (pABA)] and colorimetric detection of released pABA. Galactosemia can also be screened for via the enzymatic measurement of galactose-1-phosphate uridyltransferase (GALT) activity, which screens for the most severe form of galactosemia.

Although the aforementioned GALT assay misses galactosemia due to galactokinase (GALK) deficiency and uridine diphosphate-galactose 4′-epimerase deficiencies, these forms of galactosemia are less common than GALT deficiency and, in the case of GALK deficiency, are not associated with risk of systemic illness and death in the first few months of life. However, these other mutations should still be considered in individuals with symptoms of galactosemia. More detailed discussion of many inborn errors of metabolism can be found in Chapter 48, Inborn errors of metabolism and newborn screening.

Cystic fibrosis

Cystic fibrosis (CF) is an autosomal recessive condition caused by mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene that ultimately affects ion transport across cellular membranes in the pancreas, lung, and sweat glands. There is a broad range of phenotypes that complicate diagnosis. Furthermore, over 1500 variants have been reported in the *CFTR* gene, but the functional importance of many of these variants is unclear. The most common mutation, deltaF508 (HGVS nomenclature: p. F508del), accounts for approximately two-thirds of all mutations worldwide. Newborn screening approaches include measurement of IRT and/or mutational analysis of the *CFTR* gene. Guidelines for patient diagnosis and management by the Cystic Fibrosis Foundation rely on a combination of newborn screening results, clinical features, sweat chloride testing, and genetic analysis [17].

Newborns that screen positive for CF need to be confirmed by an alternative method, either sweat chloride testing or molecular testing. Sweat chloride testing involves three parts: (1) sweat stimulation; (2) sweat collection; and (3) sweat analysis. Sweat stimulation is accomplished by introducing a stimulant (pilocarpine) into the sweat glands with a small electrical current (1.5 mA). The stimulation is painless and only takes a few minutes to complete. Sweat is collected most commonly with a disposable plastic sweat collector that is placed on the stimulated skin area, and uses hydrostatic pressure to force the sweat into the collector. The collector contains a soluble blue dye for visual confirmation that sufficient sweat is collected. Paper or gauze methods can also be used for sweat collection. Finally, the chloride in the sweat is determined using coulometric titration on a chloridometer. Sweat chloride concentrations of >60 mEq/L are considered diagnostic of CF. The most common challenges encountered with sweat testing in the newborn population are related to specimen collection (i.e., quantity not sufficient, QNS). Sites that perform this testing are required to monitor their QNS Sufficient rate to identify problems in their training or technique.

Other conditions

Recently, newborn screening for lysosomal storage diseases (LSDs) has been made possible using multiplex enzymatic assays. One tenet of newborn screening is that early treatment improves outcome (i.e., benefits the newborn). Thus the lack of treatment for many LSDs would typically make them unsuitable for newborn screening. In addition, many of these disorders are late onset. However, early diagnosis can allow prenatal testing in future pregnancies and the potential for presymptomatic treatment with enzyme replacement therapy when it becomes available.

In 2010 the Department of Health and Human Services added SCID syndrome to the recommended uniform screening list for all states. This utilizes a unique T-cell receptor excision circle (TREC) assay, which detects T-cell lymphopenia in dried blood spots. Individuals with SCID have low or undetectable TRECs, which are a by-product of the maturation of thymocytes [18]. Most individuals with SCID can be successfully treated with a hematopoietic cell transplantation.

Follow-up

Some conditions are associated with a few common mutations, and DNA testing may be included as a second level of testing for those infants whose primary test result is consistent with the disorder. DNA testing is only definitive if the mutation on both alleles is identified. When DNA testing is not possible or does not identify both mutations, a disorder that is suggested by newborn screening results must be excluded or confirmed by any of the tests used to diagnose the disorder in a symptomatic individual. The American College of Medical Genetics has provided recommendations for clinical and laboratory follow-up for each of the conditions detectable by newborn screening (http://www.acmg.net).

Lead

According to the Centers for Disease Control (CDC), lead poisoning is the most preventable environmental exposure in children, despite the fact that approximately 500,000 children in the United States have an elevated blood lead level. Lead-contaminated dust is the primary source of lead exposure in childhood, and African−American and Hispanic children are at higher risk for exposure than Caucasian children [19]. In overt acute lead poisoning, generally associated with blood lead levels of >70 μg/dL, children may present with the following symptoms: encephalopathy with ataxia, developmental regression, poor coordination, apathy, vomiting, seizures, altered mental state, and/or coma. Physical exam may show diminished deep tendon reflexes, increased intracranial

pressure, facial nerve palsy, and/or papilledema. Milder presentations include nonspecific symptoms such as anorexia, constipation, abdominal pain, irritability, and/or altered behavior. Lower-level, chronic lead exposure (15−70 μg/dL) is associated with slowed reactions and anemia. More subtle symptoms may be missed and wrongly attributed to "normal childhood behavior." These include poor attention span, behavioral problems, and irritability. In addition, these children may present with seizures without encephalopathy, developmental delay, colicky abdominal pain, myalgia, and vomiting [20,21].

The symptoms of lead poisoning in children are nonspecific and are very subtle for blood lead levels of <25 μg/dL. A 2012 report from the CDC on chronic low-level lead exposure reported decrements in intelligence quotient (IQ) and academic achievement at blood lead levels of <10 μg/dL. Studies have shown a decrease of approximately five IQ points for every 10 μg/dL increase in blood lead levels. However, in low-range lead poisoning (1−10 μg/dL), IQ dropped by seven points, suggesting that the rate of IQ loss is greatest at lower levels [22]. Based on these findings, the CDC Advisory Committee recommended that the lowest limit of acceptability for lead be <5 μg/dL [23].

The wide range of symptoms, particularly in children with low-to-mid-level lead exposure, has prompted the American Academy of Pediatrics to recommend regular lead screening for patients deemed at risk for lead exposure [24,25]. In the 1970s and 1980s, patients were screened for elevated lead levels using free erythrocyte protoporphyrin or zinc protoporphyrin testing. Lead interferes with the production of heme, which normally consists of iron and protoporphyrin. Lead inhibits the insertion of iron into protoporphyrin, leading to a buildup of protoporphyrin, which combines nonenzymatically with zinc, to produce the fluorescent compound zinc protoporphyrin. Zinc protoporphyrin can be measured in a drop of blood in specially designed instruments called hematofluorometers. Unfortunately, zinc protoporphyrin does not increase until blood lead levels are ∼25 μg/dL. With findings that lead levels of less than 10 μg/dL can cause decreases in IQ, this test is no longer a viable method for screening patients for lead exposure.

Preferred methods for measurement of lead in blood include graphite furnace atomic absorption, anodic stripping voltammetry, and inductively coupled plasma-mass spectrometry. The capital cost for this testing is high, but the cost per test is low as reagent requirements are minimal. The LeadCare point-of-care instrument, which is based on anodic stripping voltammetry, can perform testing at clinics, and its results are reliable. However, the LeadCare Ultra analyzer has been shown to have a positive bias in low-level samples and a negative bias in high samples compared with inductively coupled plasma-mass spectrometry. Yet, around the cutoff of 5 μg/dL, no significant bias has been noted, and the analyzer is therefore considered suitable as a screening approach [26]. The capital cost for this test is low, but the individual cost is high compared with the other two methods. Another consideration with the use of point-of-care testing for lead measurements is the susceptibility for contamination in capillary collections. Elevated results by point of care need to be confirmed with a venous lead blood test [27].

Treatment

Chelation therapy is indicated in children when blood lead levels are >25 μg/dL. Oral chelation should be used for patients with blood lead levels between 45 and 69 μg/dL, and may be accomplished with succimer, an analog of dimercaprol, a compound that chelates heavy metals. Succimer binds lead, and the resulting water-soluble compound is excreted in the urine. The efficacy of this drug is modest for patients with mild to moderately elevated blood lead levels, and some studies have shown little difference compared with placebo [28]. In general, lead chelation only removes about 1%−2% of total body lead content, as chelators cannot access lead stored in some organs, such as the brain [20].

Since the late 1990s, there has been a national effort to decrease childhood lead poisoning in the United States. This program has consisted of identification of housing that may have lead paint and education of the public and physicians to the dangers of lead poisoning. The removal of lead from gasoline and paint beginning in the mid-1970s has decreased the amount of lead in the environment.

The CDC tracks cases of lead poisoning, and in 1998 reported that, in the United States, approximately 7% of children tested had a confirmed blood lead level of >10 μg/dL. That statistic dropped to <1% by 2011, suggesting that cleanup efforts, public education, and awareness have worked to decrease lead exposure in children.

Transient hyperphosphatasemia

Transient hyperphosphatasemia (TH) is a temporary condition in children under 5 years old, in which serum alkaline phosphatase (ALP) activity is elevated 3−20 times the upper reference range with no clinical indications for the elevation. TH is estimated to occur in 2%−5% of healthy children [29,30]. Patients are usually asymptomatic, with unremarkable history, physical exam and laboratory results. Some patients have had a mild viral condition in the recent past. Isoenzyme analysis is nonspecific, with both bone and liver isoenzymes usually elevated. The hyperphosphatasemia will usually resolve within 1−20 weeks without treatment [29,31]. Although clinically benign, it is important for this condition to be

recognized to prevent undue follow-ups, such as bone scans, liver biopsies, or other workups, to try and identify the cause of the elevation. It is recommended that, in patients under 5 years of age, who have elevations in both bone and liver ALP isoenzymes, and with otherwise normal lab results for liver or bone disease, an isolated elevation in ALP be attributed to TH. In these cases, ALP should be repeated in 4 months to ensure ALP has returned to normal [32].

Potassium and hemolysis

The utility of measuring potassium in pediatrics is similar to that of adult populations. Due to the critical nature of potassium for proper cardiac function and normal physiologic homeostasis, especially of the muscle and nervous systems, it is measured in serum/plasma as part of a basic metabolic panel or may be ordered on its own. For patients at high risk of potassium and other electrolyte imbalances, potassium is followed very closely [33]. Such high-risk patients include those with renal disease, diabetic ketoacidosis, and tumor lysis syndrome following chemotherapy for leukemia. Given the critical nature of this analyte, it is very important that results for potassium be accurate. Since red cells have an intracellular potassium concentration about 20 times higher than plasma, even small degrees of hemolysis can increase the potassium content in a blood sample. In newborns and infants, capillary heel stick is the recommended method of blood collection, and capillary finger stick is often used with older children (see Chapter 1, Preanalytical variation). Although these methods are less painful and invasive than venipuncture, they can cause high rates of hemolysis. Consequently, potassium results (and other hemolysis-affected tests including aspartate aminotransferase, iron, lactate dehydrogenase, haptoglobin, phosphate, and bilirubin) are often inaccurate in pediatrics and must be interpreted carefully [34].

In serum and plasma samples, the degree of hemolysis is easily detected by detecting visually or spectrophotometrically an increase in red-colored hemoglobin. For these sample types, laboratories should have established guidelines to restrict reporting of specific results on hemolyzed samples. Some have proposed the use of correction factors to estimate the true potassium concentration in hemolyzed samples. For example, Mansour et al. derived two correction factors: 0.51 or 0.40 mEq/L increased potassium for every 0.1 g/dL increase in plasma hemoglobin. However, most studies have found such factors to be insufficient for clinical implementation [35]. Potassium may be tested in whole blood using point-of-care instruments. In this situation, hemolysis cannot be identified, and results must be interpreted with caution. It is wise to repeat and verify an unexpected high potassium result from a whole blood sample before making any treatment decisions. Likewise, a normal point-of-care potassium value in a patient who was expected to be hypokalemic should also alert the clinician that hemolysis may have falsely elevated the result [36,37].

Hemolysis and plasma hemoglobin

In addition to using visual means to assess the degree of hemolysis in a plasma or serum sample, plasma hemoglobin can be measured directly with a spectrophotometer using several hemoglobin-specific wavelengths or using chemical reactions to cause a color change, which can also be measured spectrophotometrically [38]. In this way, the amount of hemoglobin in a plasma sample can be quantitated. While this is not routinely used to determine the degree of hemolysis caused by sample collection, it is used to follow hemolysis from a transfusion reaction, or due to mechanical fragmentation of red cells. Extracorporeal membrane oxygenation (ECMO) is a cause of red cell fragmentation and hemolysis seen more commonly in pediatrics than in adult practice. ECMO is an effective means of providing emergency cardiac and pulmonary aid. The process of ECMO requires circulating blood from the patient through a series of pumps, which may mechanically lyse red cells, causing an increase in free circulating hemoglobin, which puts children at risk for renal impairment and death. High plasma hemoglobin may indicate the presence of a clot in the pumps, prompting modification of anticoagulation and maintenance of the ECMO circuit [39].

Lipids

Some degree of cholesterol and lipid testing in children has been recommended for decades, as this may identify early-onset genetic lipid disorders, as well as recognizing obesity-related dyslipidemias. In 2011 the National Heart, Lung, and Blood Institute (NHLBI) published findings from an expert panel, recommending universal cholesterol screening in certain age groups. The goal of this screening was primarily to identify dyslipidemias early in life, to initiate treatment, and ultimately to decrease the risk for developing cardiovascular disease. The screening protocol recommended by the NHLBI is summarized in Fig. 29.3. The guidelines endorse universal screening in the 9–11-year-old and the 17–21-year-old age groups, with screening in other age groups limited to children with certain risk factors [40]. Despite the adoption of these recommendations by the American Association of Pediatrics, this universal screening strategy is not without controversy, particularly due to the low prevalence of disease in that population, and provider adherence to the guidelines has been low [41]. The NHLBI guidelines also published new, pediatric-specific target ranges for lipid testing, which has been more widely adopted by laboratories.

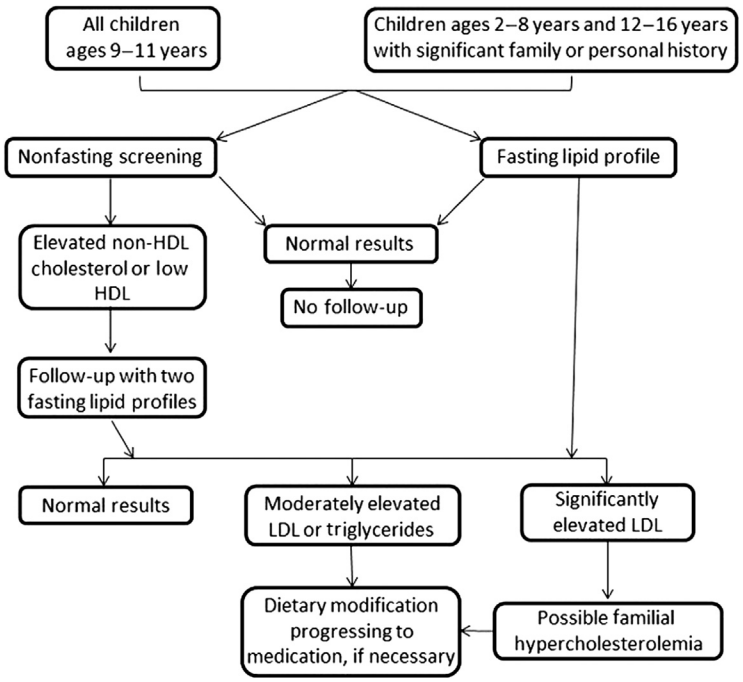


FIGURE 29.3 Algorithm summarizing the guidelines for testing lipids in children.

Bilirubin

Bilirubin is an important analyte in pediatric clinical chemistry. Approximately 60% of neonates admitted to well-baby nurseries in the United States develop hyperbilirubinemia in the first week of life [42]. Normally, bilirubin is conjugated to water-soluble glucuronide moieties, is cleared through renal and biliary mechanisms, and is nontoxic. Unconjugated bilirubin is lipid-soluble and readily crosses cell membranes, including the blood−brain barrier. Accumulation of unconjugated bilirubin in the brain is toxic, and can cause acute bilirubin encephalopathy, which may progress to kernicterus. Kernicterus is chronic and permanent bilirubin-induced brain damage. Neonates are at risk for hyperbilirubinemia, because (1) they produce increased amounts of bilirubin at birth; (2) they have low expression of glucuronosyl transferases, the enzymes responsible for glucuronide conjugation; and (3) they have high rates of red blood cell turnover. Kernicterus is a rare yet completely preventable condition. Unconjugated bilirubinemia is treatable with phototherapy or exchange transfusion in extreme cases. Effective treatment of hyperbilirubinemia relies on appropriate and accurate measurement of bilirubin.

Universally, neonates are screened for hyperbilirubinemia with total bilirubin measurement to assess risk for developing kernicterus [43]. The total bilirubin concentration is compared with the infant's age in hours. Risk assessment for hyperbilirubinemia is based on an age-specific nomogram. The nomogram is not meant to be used in premature infants or infants with significant conditions requiring neonatal intensive care, as the risk of kernicterus is higher at lower bilirubin concentrations [43]. Furthermore, the nomogram was derived from diazo methods for bilirubin and does not necessarily apply across all platforms for bilirubin measurement. As described in a study by Greene et al. [44], instruments with a positive bias for total bilirubin will result in the unnecessary use of phototherapy and those with a negative bias may miss at-risk patients. Harmonization of bilirubin measurement remains a significant challenge in pediatric clinical chemistry.

Transcutaneous bilirubin (TcB) refers to the spectrophotometric measurement of bilirubin through the skin and is measured by a handheld meter. It has become popular in inpatient units and some outpatient clinics, because it is noninvasive, fast, relatively inexpensive, and can successfully identify hyperbilirubinemia in neonates with a variety of skin tones. Hour-specific TcB reference nomograms have been developed for two commercially available, FDA cleared instruments: the BiliCheck (Phillips/Children's Medical Ventures, Monroeville, PA, United States) and JM-103 (Draeger Medical Systems, Telford, PA, Unites States) [16]. Correlation between TcB and total serum bilirubin has been demonstrated with correlation coefficients ranging between 0.71 and 0.93. TcB measurement is dependent on the skin location of the measurement. TcB measurements taken from the heel, back, or thigh do not correlate as well as measurements from the forehead or sternum. Melanin content and dermal thickness also

affect TcB measurements. Bruised, birth-marked, or hairy skin should not be used. Thus appropriate training in the use of this point-of-care instrument is critically important, and vigorous quality assurance is required [43]. TcB is a screening tool; if TcB indicates clinically significant hyperbilirubinemia, this must be confirmed by serum total bilirubin measurements. In addition, TcB only estimates total bilirubin; TcB cannot distinguish between conjugated and unconjugated fractions. Total serum bilirubin remains the gold standard for assessment of neonatal jaundice.

Conditions other than hyperbilirubinemia of the newborn that give rise to bilirubin abnormalities in infancy or childhood include hepatic biliary obstruction and the inherited disorders of bilirubin metabolism: Gilbert, Crigler–Najjar, Rotor, and Dubin–Johnson syndromes (Fig. 29.4). Of the four disorders of bilirubin metabolism, Gilbert syndrome is the most common. Gilbert and Crigler–Najjar syndromes are due to defects in UDP-glucuronosyltransferase (UGT1A1), the enzyme responsible for bilirubin conjugation, and are characterized by unconjugated hyperbilirubinemia. Gilbert syndrome is associated with 10%–35% enzyme activity and mild hyperbilirubinemia (total bilirubin values rarely exceed 6 mg/dL); infants with Gilbert syndrome are treated with phenobarbitone, which induces enzyme activity. There are two types of Crigler–Najjar syndrome. Type 1 is the most severe and is characterized by a near complete lack of UGT1A1 enzyme activity and with severe unconjugated bilirubinemia (>20 mg/dL) in the first few days of life, with no other biochemical evidence of abnormal liver function. Death from neurological sequelae due to kernicterus typically occurs by 18 months in the absence of intensive treatment such as liver transplantation. Type 2 is caused by a partial enzyme deficiency (<10%) and in contrast to Type 1, leads to less severe unconjugated hyperbilirubinemia (6–20 mg/dL). Clinical manifestations may occur if the disease is undetected for many years. Dubin–Johnson syndrome is caused by defects in a transporter gene, *ABCC2* (codes MRP2 protein), resulting in defective excretion of conjugated bilirubin. Biochemically, it is characterized by mixed hyperbilirubinemia. Rotor syndrome presents very similarly to Dubin–Johnson, but the genetic basis for Rotor syndrome is unknown, though there is some evidence that it is the result of defective uptake of unconjugated bilirubin into the hepatocyte [45]. Biliary atresia, caused by hepatic biliary obstruction, is another source of hyperbilirubinemia that presents in the neonatal period. With biliary atresia, hyperbilirubinemia is usually accompanied by elevated liver enzymes.

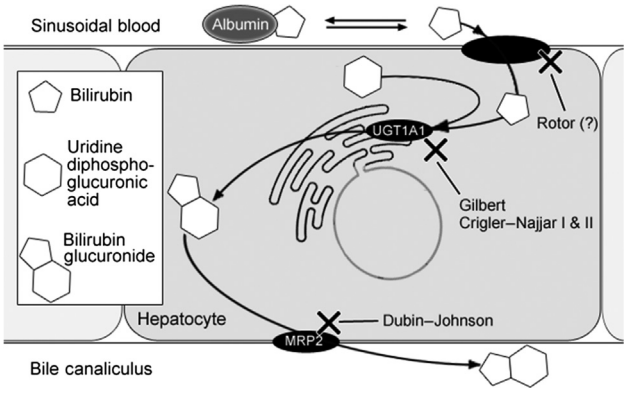


FIGURE 29.4 Disorders of bilirubin transport and conjugation.

Implications for testing in the pediatric population

Laboratory testing in the pediatric population is complicated by the fact that the results can have unintended consequences, ultimately affecting the entire family. This is due to a constellation of three factors: (1) all care providers are mandatory reporters of suspected child abuse; (2) the pediatric population is vulnerable and reliant on others for care and protection; and (3) results reflecting genetically inherited conditions can have important implications for relations, as well as unintended results for adult-onset conditions in the pediatric patient.

For the first two factors, it is important to be aware that care providers are not always able to interpret the lab results with the appropriate filters (e.g., this is a screening test, or has other known limitations). Laboratory testing necessary to diagnose certain inherited conditions may be missed altogether, leading to the wrong conclusion to be drawn (e.g., a urine drug screen may not detect a small but significant pediatric drug exposure, and bruising from bleeding disorders could be mistaken for child abuse). For these reasons, we recommend that pediatric laboratory professionals maintain familiarity with the limitations of lab tests most at risk for misinterpretation and underutilization, and intervene when necessary (e.g., urine drug screens and coagulation workups). Finally, there are guidelines recommending the appropriate approach to genetic screening and testing in children [46]. Guidelines include ensuring appropriate counseling measures for the family before testing is pursued, and deferring predictive genetic testing for adult-onset conditions when medically appropriate.

Transitioning to adult care

The final consideration in pediatric laboratory medicine is the ultimate goal of transitioning patients to adult care. For healthy children, this simple process may require changing primary care providers. For children living with chronic conditions, however, the most successful transitions take place over several years starting at 18 years of age, and ending pediatric care around 22 years old. Chronic conditions

include solid organ and bone marrow transplants, diabetes, and some inherited conditions such as muscular dystrophy and sickle cell disease. These conditions are often lab-intensive, and the transition involves obtaining blood samples at adult collection facilities and analysis, which may take place using different methodologies. Patients with some rare pediatric disorders will continue their care with pediatric specialists, because the disorders are either too rare or the prognosis of such conditions prohibit many of the affected patients from reaching adulthood.

References

[1] S.R.C. Howie, Blood sample volumes in child health research: review of safe limits, Bull. World Health Organ. 89 (2011) 46–53.

[2] S.L. Valentine, S.T. Bateman, Identifying factors to minimize phlebotomy-induced blood loss in the pediatric intensive care unit, Pediatr. Crit. Care Med. 13 (2012) 22–27.

[3] V. Bush The hemolyzed specimen: causes, effects, and reduction. BD LabNotes, 2003.

[4] C.A. Burtis, Sample evaporation and its impact on the operating performance of an automated selective-access analytical system, Clin. Chem. 36 (1990) 544–546.

[5] C.A. Burtis, J.M. Begovich, J.S. Watson, Factors influencing evaporation from sample cups and assessment of their effect on analytical error, Clin. Chem. 21 (1975) 1907–1917.

[6] WHO Guidelines on Drawing Blood: Best Practices in Phlebotomy. Geneva, 2010.

[7] P.M. Jones, Pediatric clinical biochemistry: why is it different? in: D.J. Dietzen, M.J. Bennett, E.C. Wong (Eds.), Biochemical and Molecular Basis of Pediatric Disease, AACC Press, Washington, DC, 2010. p. 660; xx.

[8] P. Greenland, N.L. Bowley, B. Meiklejohn, et al., Blood cholesterol concentration: fingerstick plasma vs venous serum sampling, Clin. Chem. 36 (1990) 628–630.

[9] M.M. Bond, R.R. Richards-Kortum, Drop-to-drop variation in the cellular components of fingerprick blood: implications for point-of-care diagnostic development, Am. J. Clin. Pathol. 144 (2015) 885–894.

[10] D. Montgomery, C. Plate, S.C. Alder, et al., Testing for fetal exposure to illicit drugs using umbilical cord tissue vs meconium, J. Perinatol. 26 (2006) 11–14.

[11] F. Ceriotti, Establishing pediatric reference intervals: a challenging task, Clin. Chem. 58 (2012) 808–810.

[12] W.L. Roberts, Hemoglobin constant spring can interfere with glycated hemoglobin measurements by boronate affinity chromatography, Clin. Chem. 53 (2007) 142–143.

[13] D.J. Dietzen, M.J. Bennett, S.F. Lo, et al., Dried blood spot reference intervals for steroids and amino acids in a neonatal cohort of the National Children's Study, Clin. Chem. 62 (2016) 1658–1667.

[14] W.H.W.R. Hannon, B. Davin, P. Fernhoff, T. Halonen, M. Lavochkin Blood collection on filter paper for newborn screening programs; Approved guideline. Clinical and Laboratory Standards Institute 2007; Document LA4-A5.

[15] P.L. Hall, G. Marquardt, D.M. McHugh, et al., Postanalytical tools improve performance of newborn screening by tandem mass spectrometry, Genet. Med. 16 (2014) 889–895.

[16] B.R.P. Wilcken, D. Matern, Newborn screening for inborn errors of metabolism, in: J.W.J. Saudubray, G. van den Berghe (Eds.), Inborn Metabolic Diseases, Springer, New York, 2012.

[17] P.M. Farrell, B.J. Rosenstein, T.B. White, et al., Guidelines for diagnosis of cystic fibrosis in newborns through older adults: Cystic Fibrosis Foundation consensus report, J. Pediatr. 153 (2008) S4–S14.

[18] J.M. Puck, Neonatal screening for severe combined immunodeficiency, Curr. Opin. Pediatr. 23 (2011) 667–673.

[19] B.P. Lanphear, R. Hornung, M. Ho, et al., Environmental lead exposure during early childhood, J. Pediatr. 140 (2002) 40–47.

[20] F.M. Henretig, Lead, in: L.S. Nelson, N.A. Lewis, M. Howland, R. S. Hoffman, L.R. Goldfrank, N.E. Flomenbaum (Eds.), Goldfrank's Toxicologic Emergencies, McGraw-Hill, New York, NY, 2011.

[21] H. Abadin, A. Ashizawa, Y.W. Stevens, et al. Toxicological Profile for Lead. Atlanta (GA), 2007.

[22] R.L. Canfield, C.R. Henderson, D.A. Cory-Slechta, et al., Intellectual impairment in children with blood lead concentrations below 10 μg per deciliter, N. Engl. J. Med. 348 (2003) 1517–1526.

[23] K.S. Betts, CDC updates guidelines for children's lead exposure, Environ. Health Perspect. 120 (2012). a268-a268.

[24] Screening for elevated blood lead levels. American Academy of Pediatrics Committee on Environmental Health. Pediatrics 1998; 101:1072-1078.

[25] Practice CO, Ambulatory Medicine BFPSW, 2014 recommendations for pediatric preventive health care, Pediatrics 133 (2014) 568–570.

[26] L.M. Johnson, M. Lynch, C. Johnson, et al., Evaluation of LeadCare Ultra(R) as an initial screen for elevated blood lead levels, Clin. Biochem. 66 (2019) 95–99.

[27] Guidelines for measuring lead in blood using point of care instruments: Advisory Committee on Childhood Lead Poisoning Prevention of the Centers for Disease Control and Prevention. 2013.

[28] M.E. Mortensen, P.D. Walson, Chelation therapy for childhood lead poisoning. The changing scene in the 1990s, Clin. Pediatr. (Phila.) 32 (1993) 284–291.

[29] S.Y. Huh, H.A. Feldman, J.E. Cox, et al., Prevalence of transient hyperphosphatasemia among healthy infants and toddlers, Pediatrics 124 (2009) 703–709.

[30] P. Ridefelt, J. Gustafsson, M. Aldrimer, et al., Alkaline phosphatase in healthy children: reference intervals and prevalence of elevated levels, Horm. Res. Paediatr. 82 (2014) 399–404.

[31] G. Gualco, S.A. Lava, L. Garzoni, et al., Transient benign hyperphophatasemia, J. Pediatr. Gastroenterol. Nutr. 57 (2013) 167–171.

[32] J.L. Otero, R.P. Gonzalez-Peralta, J.M. Andres, et al., Elevated alkaline phosphatase in children: an algorithm to determine when a "wait and see" approach is optimal, Clin. Med. Insights Pediatr. 5 (2011) 15–18.

[33] S.A. Lanham-New, H. Lambert, L. Frassetto, Potassium, Adv. Nutr. 3 (2012) 820–821.

[34] M. Oostendorp, W.W. van Solinge, H. Kemperman, Potassium but not lactate dehydrogenase elevation due to in vitro hemolysis is higher in capillary than in venous blood samples, Arch. Pathol. Lab. Med. 136 (2012) 1262–1265.

[35] M.M. Mansour, H.M. Azzazy, S.C. Kazmierczak, Correction factors for estimating potassium concentrations in samples with in vitro hemolysis: a detriment to patient safety, Arch. Pathol. Lab. Med. 133 (2009) 960–966.

[36] S.C. Kazmierczak, A.F. Robertson, K.P. Briley, Comparison of hemolysis in blood samples collected using an automatic incision device and a manual lance, Arch. Pediatr. Adolesc. Med. 156 (2002) 1072–1074.

[37] A. Algeciras-Schimnich, W.J. Cook, T.C. Milz, et al., Evaluation of hemoglobin interference in capillary heel-stick samples collected for determination of neonatal bilirubin, Clin. Biochem. 40 (2007) 1311–1316.

[38] V.F. Fairbanks, S.C. Ziesmer, P.C. O'Brien, Methods for measuring plasma hemoglobin in micromolar concentration compared, Clin. Chem. 38 (1992) 132–140.

[39] W. Butt, M. Heard, G.J. Peek, Clinical management of the extracorporeal membrane oxygenation circuit, Pediatr. Crit. Care Med. 14 (2013) S13–S19.

[40] Lipids and lipoproteins. Expert Panel on Integrated Guidelines for Cardiovascular Health and Risk Reduction in Children and Adolescents: NIH, 2012.

[41] A.J. Smith, E.L. Turner, S. Kinra, Universal cholesterol screening in childhood: a systematic review, Acad. Pediatr. 16 (2016) 716–725.

[42] C.C.E. Revenis MaW, Disorders in the neonatal period, in: D.J. Dietzen, M.J. Bennett, E.C. Wong (Eds.), Biochemical and Molecular Basis of Pediatric Disease, AACC Press, Washington, DC, 2010. p. 660; xx.

[43] American Academy of Pediatrics Subcommittee on H, Management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation, Pediatrics 114 (2004) 297–316.

[44] D.N. Greene, J. Liang, D.T. Holmes, et al., Neonatal total bilirubin measurements: still room for harmonization, Clin. Biochem. 47 (2014) 1112–1115.

[45] S. Erlinger, I.M. Arias, D. Dhumeaux, Inherited disorders of bilirubin transport and conjugation: new insights into molecular mechanisms and consequences, Gastroenterology 146 (2014) 1625–1638.

[46] Committee On B, Committee On Genetics AND, American College Of Medical Genetics AND, et al., Ethical and policy issues in genetic testing and screening of children, Pediatrics 131 (2013) 620–622.

Self-assessment questions

1. A 48-h-old term infant has a serum bilirubin of 12. According to the bilirubin nomogram, this puts the infant in which risk category?
   - **a.** Low risk
   - **b.** High-intermediate risk
   - **c.** High risk
   - **d.** The nomogram cannot be used to predict risk in term infants.
2. Which of the following conditions was the first to be mandated for newborn screening?
   - **a.** Cystic fibrosis
   - **b.** Phenylketonuria
   - **c.** Severe combined immunodeficiency
   - **d.** Kernicterus
3. Which of the following contribute the most to the minimum volume required for an assay?
   - **a.** Patient total blood volume
   - **b.** Analytical volume
   - **c.** Dead volume
   - **d.** The need for duplicate testing
4. A newborn screen is positive for cystic fibrosis. Follow-up sweat chloride testing was resulted as 40 mmol/L. This is interpreted as _____.
   - **a.** Negative for cystic fibrosis
   - **b.** Negative for cystic fibrosis, but positive for pulmonary atresia
   - **c.** Positive for cystic fibrosis
   - **d.** Indeterminate
5. Challenges with establish reference intervals in the pediatric population include _____.
   - **a.** Parental willingness to participate with healthy children
   - **b.** Federal regulations
   - **c.** Changes in analyte concentrations during development
   - **d.** Previously established reference intervals were with outdated methodology.
   - **e.** A, C, and D

Answers

1. b
2. b
3. c
4. a
5. e

Chapter 30

Biomarkers for coronary artery disease and heart failure

Peter A. Kavsak[1,2] and Alan H.B. Wu[3]

[1]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, ON, Canada, [2]Juravinski Hospital and Cancer Centre, Hamilton Health Sciences, Hamilton, ON, Canada, [3]Department of Laboratory Medicine, University of California, San Francisco, San Francisco General Hospital, San Francisco, CA, United States

Learning objectives

At the end of this chapter, the reader will be able to:

- Describe the pathophysiology of acute coronary artery disease and heart failure (HF).
- Compare the diagnostic criteria from the World Health Organization with that of the European Society/American College of Cardiology to that of the Universal Myocardial Infarction definitions, with a focus on high-sensitivity cardiac troponin assays.
- Explain how cardiac troponin and B-type natriuretic peptides have become the standard blood biomarkers for acute coronary syndrome and HF, respectively.
- List novel biomarkers and how they may complement current biomarkers.
- Describe the immunoassay measurement technologies for cardiac biomarkers.

Introduction

Cardiovascular disease (CVD) continues to be an important cause of death throughout the world. In the United States (US), from 2001 to 2011, the death rate due to CVD declined by approximately 30%; however, CVD was still attributed as the cause of death in nearly one-third of all deaths in 2011 [1]. Broken down further into the major cardiovascular conditions/diseases, coronary heart disease caused ~1 in 7 deaths, heart failure (HF) ~1 in 9 deaths, and stroke causing ~1 in 20 deaths in the US [1]. Globally, CVD is the number one cause of death, with both age and population growth acting as contributors to the increased frequency; nearly one-third of deaths globally were attributed to CVD in 2013, with ischemic heart disease and stroke being predominant causes in most countries [2].

Diagnostic tests, in particular blood-based biomarkers, have played an integral role in CVD for over 60 years, especially for the diagnosis of acute myocardial infarction (AMI), where initially the enzymatic activity of serum aspartate aminotransferase (AST) was demonstrated to be increased in patients with AMI [3,4]. Since the turn of the millennium, however, cardiac troponin has taken a prominent role and is considered the most optimal blood-based biomarker to be used in the diagnosis of AMI. Outside of the diagnosis of AMI, cardiac troponin is also used for risk stratification for future adverse cardiac events, and in therapeutic selection such as with platelet inhibitors and acute intervention such as with percutaneous intervention (PCI; an example being balloon angioplasty). The prevalence of HF in the population has increased, with projections in the US suggesting close to a 50% increase in the HF prevalence from 2012 to 2030, resulting in more than 8 million adults living with HF [1]. Fortunately, the blood-based biomarkers B-type natriuretic peptide (BNP) and N-terminal-proBNP (NT-proBNP) are available and may be used for the diagnosis of HF and for the management of patients with HF. Additional/novel biomarkers have been developed and more are being evaluated to aid in the diagnosis, prognosis, and management of the various cardiovascular conditions/diseases, with the development and validation of biomarkers for ischemic heart disease and HF garnering the most attention and effort in this area.

Pathophysiology of acute coronary syndrome and myocardial infarction classification

Coronary artery disease (CAD) usually has a progressive pathophysiology that typically begins with atherosclerosis, the narrowing of coronary arteries by the deposition of

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00030-2

lipid-filled plaques, to vulnerable plaques, and finally to acute coronary syndrome (ACS), a collective term that refers to acute rupture, fissuring, or erosion of vulnerable plaques [5,6]. Fig. 30.1 illustrates the various stages of atherosclerosis. There are no symptoms associated with the initial narrowing of the coronary arteries (stages 2–4). However, with increasing coronary artery stenosis (stage 5), symptoms of chest pain begin first in conjunction with physical exertion (stable angina). These individuals have plaques that are weakened and prone to rupture. The onset of chest pain without exertion signifies the onset of ACS, as seen with the rupture of the coronary artery plaque (stage 6). The release and exposure of the plaque contents ("gruel") to circulating blood cause activation of platelets and formation of a thrombus or distal platelet emboli, which further decreases myocardial blood flow and sets the stage for tissue necrosis. In the historical World Health Organization (WHO, 1979) definition of AMI, a complete artery blockage by the clot was termed a "transmural" AMI, while a subocclusive clot was termed a "subendocardial" infarct. Patients with minimal or no release of biomarkers were termed "unstable angina" [7]. In the latest myocardial infarction (MI) definition, the Fourth Universal Definition of myocardial infarction (2018), the above pathophysiological process would be classified as a spontaneous MI (Type 1 MI), with unstable angina diagnosed if these patients have no biomarker elevation [8].

In addition to elevated biomarkers, patients with Type 1 MI often have electrocardiogram (ECG) changes [8–10]. The ECG finding is important for treatment strategies, and for patients with MI who develop ST elevation in two contiguous leads on the ECG, the MI can be clinically classified as ST elevation MI (STEMI); those MI patients who do not have this ECG finding are classified as non-ST elevation MI (NSTEMI) [9]. Type 1 MI includes processes that involve plaque rupture/erosion with an occlusive thrombus or with a nonocclusive thrombus [8]. In addition to the Type 1 MI classification, the Fourth Universal Definition of MI has further provided additional classification into the other four types of MI: Type 2 MI, Type 3 MI, Type 4 MI, and Type 5 MI.

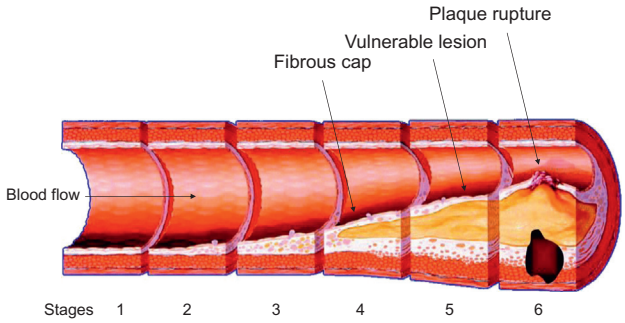


FIGURE 30.1 Pathophysiology of acute coronary syndrome. The initial four stages are not usually accompanied by symptoms. Chest pain upon exertion (stable angina) can occur in stage 5. In stage 6, there is plaque rupture causing clot formation and acute coronary syndrome. *Adapted and used with permission from F.S. Apple, A.H.B. Wu, J. Mair, J. Ravkilde, M. Panteghini, J. Tate, et al., Future biomarkers for detection of ischemia and risk stratification in acute coronary syndrome, Clin. Chem. 51 (2005) 810–24.*

Type 2 MI requires evidence of an imbalance between myocardial oxygen supply and demand, which is not due to an acute coronary atherothrombosis (i.e., not Type 1 MI) [8]. Besides clinical conditions leading to an oxygen supply/demand imbalance alone resulting in Type 2 MI, patients with nonatherosclerotic coronary dissection, vasospasm, or coronary microvascular dysfunction and those with atherosclerosis and oxygen supply/demand imbalance may result in Type 2 MI [8]. Thus context is important, (i.e., Are patients presenting with chest pain or is it secondary to another illness?). Mechanisms that may lead to Type 2 MI include hypotension/shock, severe anemia, respiratory failure, severe bradyarrhythmia, severe hypertension, sustained tachyarrhythmia, drug-induced vasospasm (e.g., amphetamine, cocaine), coronary spasm/microvascular dysfunction, coronary embolism, fixed coronary atherosclerosis, and coronary artery dissection [8]. Accordingly, Type 2 MI represents a challenge, as a condition other than CAD may contribute to the development of MI due to an oxygen supply/demand imbalance that can occur in many hospitalized patients (e.g., noncardiac surgery patients, critically ill patients, etc.) [8–11]. Type 3 MI relates to patients who suffer from cardiac death (i.e., have symptoms suggestive of myocardial ischemia) and die before blood is obtained or before increases in biomarker concentrations are evident or the MI is identified by autopsy [8]. Type 4 MI is subclassified into 4a, 4b, and 4c and all relate to PCI. Specifically, Type 4a is a PCI-related MI (i.e., $\leq$ 48 hours after index procedure), Type 4b is a stent/scaffold thrombosis associated with PCI, and Type 4c is a restenosis associated with PCI [8]. Type 5 MI is a coronary artery bypass grafting (CABG)-related MI (i.e., $\leq$48 hours after index procedure) [8].

Another important physiological process that has a role in the development of CAD and ACS is inflammation. Inflammation stimulates the recruitment of macrophages, monocytes, and foam cells to the shoulder regions of the coronary artery. These inflammatory cells release degradative enzymes such as metalloproteinases, myeloperoxidase, and collagenases that degrade the collagen-rich fibrous cap making plaques vulnerable for rupture and erosion. They may also impair release of nitric oxide, thereby inhibiting vasodilation and further precipitating the vasospasmic etiology of AMI. It should be noted that plaque rupture and erosion do not always occur on the arteries with the highest degree of stenosis. Vulnerable plaques can exist throughout the coronary artery

vasculature. An understanding of the etiologies and the classifications of MI is important for interpreting the results of the current cardiac biomarkers, in the development of new biomarkers and in monitoring and assessing new treatments [12].

Evolving definitions of myocardial infarction with an increasing focus on biomarkers

The criteria for the definition of MI have evolved over the years, in part due to the improvement in detection technologies, in particular biomarkers. In 1979 the WHO determined that two of three criteria were necessary for the diagnosis of AMI [7]. As listed in Table 30.1, these included clinical history, electrocardiographic changes, and changes in serial enzyme results. Serial enzyme results largely referred to as the enzymes creatine kinase (CK), lactate dehydrogenase (LD), and their creatinine kinase-myocardial band isoenzyme (CK-MB) and LD-1 isoenzymes. These criteria were established before the development and implementation of cardiac troponin. Under the WHO criteria, it was possible to diagnose AMI in the absence of an increase in CK-MB or LD-1 enzyme activity. Given that these tests are not as specific and analytic as sensitive as the current cardiac troponin tests, it was possible to have a patient with a mild AMI that had enzyme results that were within that enzyme's reference interval.

The development of immunoassays for CK-MB mass as well as the discovery of cardiac troponin and development of the associated immunoassay tests heralded a new era for laboratory medicine practices for the diagnosis of AMI [4,13]. Importantly, cardiac troponin is a protein (nonenzyme) biomarker that has been shown to be more analytically sensitive and specific than the existing enzyme markers. As a result, in the year 2000, a joint committee of the European Society of Cardiology (ESC)/American College of Cardiology (ACC) redefined the criteria for AMI to be predicated on documenting an increase in cardiac troponin or CK-MB mass (total CK, AST, and LD not recommended) in the context of myocardial ischemia (Table 30.1) [14]. Evidence of ischemia includes clinical history, electrocardiographic changes, angiographic evidence, and anatomic analysis (at autopsy). Therefore unlike the WHO criteria, the ESC/ACC redefinition of AMI was now predicated on an increase of a specific cardiac biomarker measured with immunoassays. Also the ESC/ACC redefinition stated that the 99th percentile upper reference limit (URL) from a healthy population should be used as the cutoff for detecting myocardial necrosis with the cardiac protein biomarkers [14].

With the 2007 Universal Definition, the preferred cardiac biomarker for detecting myocardial necrosis was stated to be cardiac troponin, with CK-MB mass as an alternative assay only if cardiac troponin was not available [15]. The 99th percentile URL was again listed as the cutoff, with an important caveat that "the use of assays that do not have independent validation of optimal precision (coefficient of variation, CV $\leq$ 10%) is not recommended." [15]. One of the important motivating factors for changing to a cardiac troponin standard (besides the myocardial tissue specificity) was the finding that patients who had minor increases in cardiac troponin concentrations were at risk of CVD progression and death, similar to those who had an AMI caused by a totally occlusive clot (i.e., high clinical sensitivity). Therefore these patients should be aggressively managed to reduce morbidity and mortality, and movement to using cardiac troponin as the standard and the 99th percentile URL as the cutoff via the 2000 and 2007 AMI definitions enabled justification of the resources needed. Under these definitions, patients who may have had minor myocardial damage as determined by low cardiac troponin concentrations and were previously diagnosed as unstable angina would now be diagnosed with NSTEMI. The impact of moving from the WHO criteria and high cardiac biomarker cutoffs to the redefinition of MI using cardiac troponin and the 99th percentile URL cutoff is an increase in the number of AMI patients. Importantly though, these newly defined patients are at increased risk of subsequent MI /death as compared with those negative by both the old WHO AMI definition and the 2000/2007 AMI definitions (Fig. 30.2) [16,17]. Roughly, a third more of unstable angina patients were now diagnosed as AMI. Fig. 30.3 is a model that illustrates the impact of the AMI redefinition (post-2000) on the number of AMI cases seen annually in the US. A similar increase is evident worldwide.

With the 2012 Third Universal definition of MI, much has remained the same insofar that the preferred biomarker remains cardiac troponin and the 99th percentile URL is still assigned as the cutoff. Analytically though, the 2012 MI definition has ceded the 10% precision requirement at the 99th percentile URL concentration and allows cardiac troponin assays imprecision up to 20% at this concentration [9]. Also the 2012 MI definition is the first to mention high-sensitivity cardiac troponin (hs-cTn) assays. Before 2012, only Roche Diagnostics' high-sensitivity cardiac troponin T (hs-cTnT) assay was clinically available globally [US Food and Drug Administration (FDA) approval in 2017 with labeling as Troponin T Gen 5 in the US] and no other high-sensitivity cardiac troponin I (hs-cTnI) assays were regulatory approved before 2012 [18]. Accordingly, due to the limited number of studies on hs-cTn assays compared

TABLE 30.1 Comparisons of myocardial infarction definitions using the epidemiological approach: the World Health Organization (1979) versus the clinical approach: European Society of Cardiology/American College of Cardiology (European Society of Cardiology/American College of Cardiology redefinition, 2000, First), Universal (2007, Second), Third Universal (2012), and Fourth Universal (2018) Definitions of Myocardial Infarction.

WHO 1979 definition (two of the following three criteria must be met)
- History is typical if severe and prolonged chest pain is present
- Unequivocal ECG changes, abnormal Q or QS waves, evolving injury lasting longer than 1 day
- Unequivocal serial enzyme changes, or initial rise and subsequent fall of levels. Change must be properly related to the particular enzyme and the delay between onset and sampling

ESC/ACC 2000 redefinition

Typical rise and gradual fall (troponin) or more rapid rise and fall (CK-MB) of biochemical markers of myocardial necrosis with at least one of the following:
- ischemic symptoms
- development of pathologic Q waves
- ECG changes indicative of ischemia (ST segment elevation or depression)
- coronary artery intervention

Pathologic findings of AMI

Universal 2007 definition

Detection of rise and/or fall of cardiac biomarkers (preferably troponin) with at least one value above the 99th percentile of the URL together with evidence of myocardial ischemia with at least one of the following:
- Symptoms of ischemia
- ECG changes indicative of new ischemia
- Development of pathological Q waves in the ECG
- Imaging evidence of new loss of viable myocardium or new regional wall motion abnormality

Criteria for MI Types 3,4, and 5

Pathologic findings of AMI

Third Universal 2012 definition
- Detection of rise and/or fall of cardiac biomarker values (preferably cTn) with at least one value above the 99th percentile upper reference limit (URL) and with at least one of the following:
  - Symptoms of ischemia
  - New or presumed new significant ST-segment-T wave changes or new left bundle branch block
  - Development of pathological Q waves in the ECG
  - Imaging evidence of new loss of viable myocardium or new regional wall motion abnormality
  - Identification of an intracoronary thrombus by angiography or autopsy
- Criteria for MI Types 3,4, and 5

Fourth Universal 2018 definition

The term acute myocardial infarction should be used when there is acute myocardial injury with clinical evidence of acute myocardial ischemia and with detection of a rise and/or fall of cardiac troponin concentrations with at least once concentration above the 99th percentile URL and at least one of the following:
- Symptoms of myocardial ischemia
- New ischemic ECG changes
- Development of pathological Q waves
- Imaging evidence of new loss of viable myocardium or new regional wall motion abnormality in a pattern consistent with an ischemic etiology
- Identification of a coronary thrombus by angiography or autopsy (not for type 2 or 3 MIs)

Postmortem demonstration of acute atherothrombosis in the artery supplying the infarcted myocardium meets criteria for Type 1 MI. Evidence of an imbalance between myocardial oxygen supply and demand unrelated to acute atherothrombosis meets criteria for type 2 MI. Cardiac death in patients with symptoms suggestive of myocardial ischemia and presumed new ischemic ECG changes before cardiac troponin concentrations become available or abnormal meets criteria for type 3 MI.

Criteria for MI Types 4 and 5

Criteria for prior or silent/unrecognized MI: abnormal Q waves; imaging evidence of loss of viable myocardium; and pathoanatomical findings of a prior MI

ACC, American College of Cardiology; *AMI*, acute myocardial infarction; CK, creatine kinase; *cTn*, cardiac troponin; *ECG*, electrocardiogram; *ESC*, European Society of Cardiology; *MI*, myocardial infarction; *URL*, upper reference limit; *WHO*, World Health Organization.

with the sensitive cardiac troponin assays at the time of writing the Third Universal Definition of MI, there was limited text on these assays for use in MI diagnosis. The following was the only statement on hs-cTn assays in the *Biomarker Detection of Myocardial Injury with Necrosis* section of the consensus document: "Sex-dependent values may be recommended for high-sensitivity troponin assays" [9]. A formal definition of hsTn and further

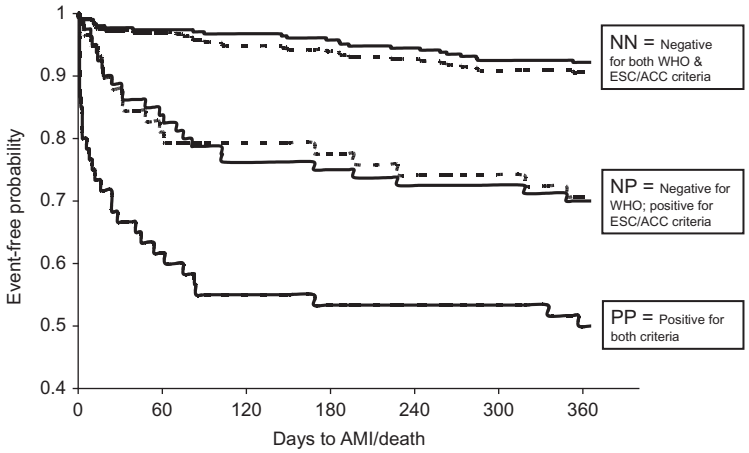


FIGURE 30.2 Kaplan–Meier curves for mortality from all causes and/or subsequent acute myocardial infarction in patients who presented to the emergency department with symptoms suggestive of cardiac ischemia. Among 448 patients stratified by two definitions of acute myocardial infarction, patients were stratified into three groups (NN—negative for both acute myocardial infarction definitions; NP—positive for the European Society of Cardiology/American College of Cardiology acute myocardial infarction definition only; and PP—Positive for both acute myocardial infarction definitions). Solid line indicates use of the 99th percentile cutoff level for cardiac troponin I in the European Society of Cardiology/American College of Cardiology criteria; dotted line indicates use of the 10% cardiovascular cutoff level for cardiac troponin I in the European Society of Cardiology/American College of Cardiology criteria. *From P.A. Kavsak, A.R. MacRae, G.E. Palomaki, A.M. Newman, D.T. Ko, V. Lustig, et al., Health outcomes categorized by current and previous definitions of acute myocardial infarction in an unselected cohort of troponin-naïve emergency department patients. Clin. Chem. 52 (2006) 2028–35 with permission.*

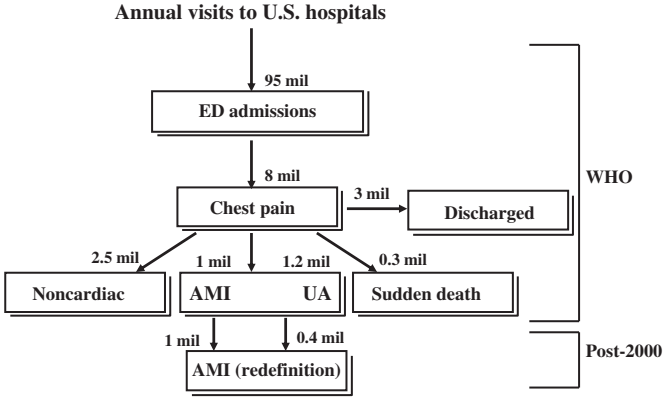


FIGURE 30.3 Emergency department visits for patients with chest pain and the incidence of myocardial infarction under the World Health Organization and the post-2000 redefinition of myocardial infarction (e.g., European Society of Cardiology/American College of Cardiology criteria etc.).

discussion of its utility will be given in the subsequent sections of the chapter.

The Fourth Universal Definition of Myocardial Infarction released in August 2018 has further fleshed out the difference between myocardial injury and MI. Specifically, myocardial injury is identified if the cardiac troponin concentration is above the 99th percentile URL, with acute injury evident if there is a rise and or fall in cardiac troponin concentrations [8]. The term AMI should be used when there is acute myocardial injury with clinical evidence of acute myocardial ischemia [8]. The preferred biomarker of myocardial injury remains cardiac troponin, with hs-cTn assays recommended for routine use. Serial measurements are required to differentiate if the myocardial injury is acute or chronic, with $\leq 20\%$ variation in concentrations indicative of stable elevated levels [8]. The imprecision at the 99th percentile concentration must be $\leq 10\%$ CV for hs-cTn assays, with imprecision between 10% and 20% acceptable at the 99th percentile for the non-hs-cTn assays (i.e., the contemporary assays) [8]. For hs-cTn assays, sex-specific 99th URLs are recommended, despite being "controversy as to whether this approach provides valuable additional information for all hs-cTn assays." [8], according to the ESC 4th Universal Definition of Myocardial Infarction. Finally, the fourth definition still recommends baseline

(0 hour) measurement with a repeated measurement 3−6 hours later, with earlier measurements possible when using hs-cTn assays [8].

Cardiac troponins T and I

Biochemistry

The troponin protein complex binds to tropomyosin and actin of the thin filament and regulates contraction within striated muscle. It consists of three proteins: troponin C (molecular weight ∼ 18 kDa) calcium binding protein; troponin I (molecular weight <30 kDa) protein that inhibits ATPase, and troponin T (molecular weight ≤40 kDa) anchoring protein that binds to tropomyosin. Cardiac troponin T (human cTnT = 298 amino acids) and I (human cTnI = 210 amino acids) isoforms from the heart are structurally distinct from the corresponding skeletal muscle isoforms. Therefore the concentration of cardiac troponin in the patient's blood is not influenced by the presence of most types of skeletal injury, making the test for cardiac troponins T and I highly specific for the heart [19]. Assays for cardiac troponin T and I are also more sensitive than CK-MB, as these subunits have a higher myocardial content (see Table 30.2). The subcellular distribution of cardiac troponin is shown in Fig. 30.4; though this and the release kinetics and forms of cardiac troponin are still an area of active research [20].

TABLE 30.2 Myocardial tissue content of cardiac markers.

Marker	Content
Cardiac troponin T	∼10.8 mg/g wet weight
Cardiac troponin I	∼4.0−6.0 mg/g wet weight
Creatine kinase-MB	∼1.4 mg/g wet weight

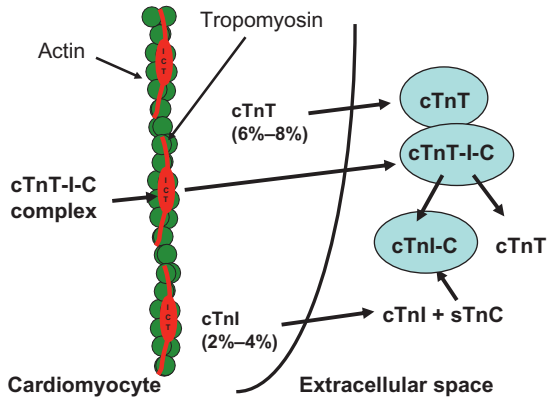


FIGURE 30.4 Subcellular distribution of cardiac troponin within the cardiomyocyte and release into the circulation following injury.

Approximately 6%−8% of cTnT and 2%−4% of cTnI are found free in the cytoplasm. Following irreversible myocardial injury, the cytosolic forms are first to appear in blood (Fig. 30.5A). This is followed by the prolonged release of the ternary complex of cardiac troponin (T−I−C) due to the breakdown of the myofibrils by degradative enzymes. The ternary complex further degrades in blood to the binary I−C complex and free cardiac troponin T, with fragments detected from both troponin T and troponin I; some of this degradation may occur in the necrotic cells from patients with AMI [21−23]. However, there are other mechanisms (i.e., ischemia and reversible injury) that may result in increased concentrations of cardiac troponin [24,25]. Specifically, both cell culture and animal models have demonstrated bleb formation in the presence of anoxia (i.e., bubbles in plasma membrane due to ischemia; see Fig. 30.5B) [24], with a large meta-analysis demonstrating that cardiac troponin (as well as other biomarkers used in the acute care setting, e.g., D-dimer and the natriuretic peptides) can increase after strenuous exercise [25].

Although there are some differences (depending on the manufacturer and assay version used), the clinical performance of cTnT and cTnI overall for aiding in the diagnosis of AMI is similar, and a clinical laboratory would only need to offer one test, not both tests, for this purpose. However, in rare circumstances, measurements of both may be helpful in specific patient conditions and during investigations of possible interferences [26,27]. To this end, for practicality, rather than having more than one cardiac troponin test available in this situation, a clinical laboratory could refer problematic samples to another laboratory running an alternate cardiac troponin assay to aid in the investigation.

Clinical utility of cardiac troponins in acute coronary syndrome

Diagnosis and risk stratification

Cardiac troponin is the most important biomarker for aiding in a diagnosis and for the risk stratification of patients presenting with symptoms suggestive of ACS. However, for an STEMI diagnosis, this can be achieved without laboratory testing, though depending on the time from pain onset and time of blood collection, cardiac troponin concentrations would be elevated in patients presently acute with this condition. Using the 99th percentile URL concentration from a healthy normal population as the decision level for cardiac troponin means that 1% of the healthy population would have a cardiac troponin concentration exceeding this cutoff. Also, until recently, manufacturer assays had a difficult time in achieving the optimal precision at the 99th percentile concentration,

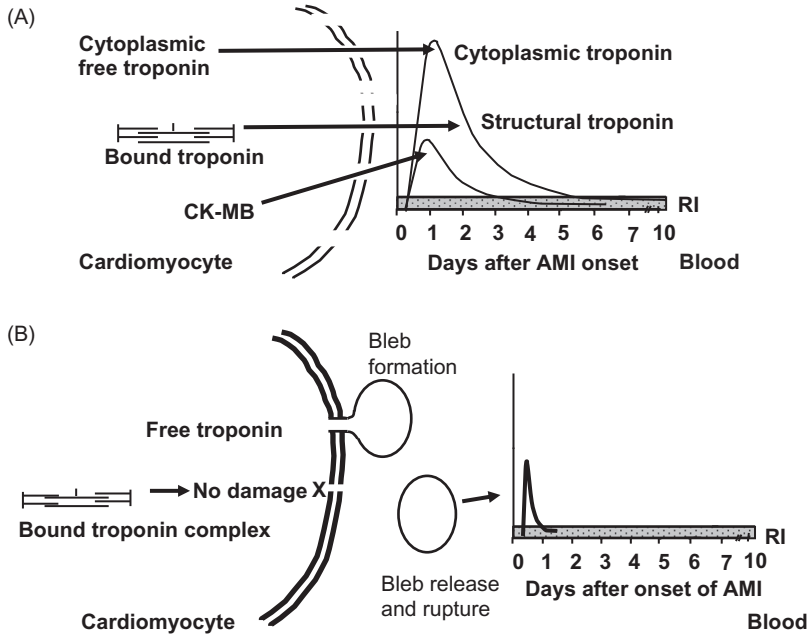


FIGURE 30.5 (A) Irreversible injury release kinetics for cardiac troponin and creatine kinase-MB following myocardial injury with cytosolic cardiac troponin first released followed by structural cardiac troponin into the blood. (B) Reversible injury with cytosolic cardiac troponin release with possible bleb formation. *RI*, Reference interval. *Adapted from P.E. Hickman, J.M. Potter, C. Aroney, G. Koerbin, E. Southcott, A.H. Wu, et al., Cardiac troponin may be released by ischemia alone, without necrosis. Clin. Chim. Acta. 411 (2010) 318–23 with permission.*

causing further analytical and clinical concerns. Accordingly, after the redefinition of MI, there were concerns on the clinical specificity of the test (i.e., the false-positives), which is largely dependent on the cutoff concentration chosen.

Historically, the cutoff concentrations for nonspecific cardiac biomarkers such as CK-MB and myoglobin (both analytes are expressed in skeletal muscle) were established to differentiate unstable angina from AMI. With the clinical implementation of the first generation cTnT and cTnI assays, the National Academy of Clinical Biochemistry (NACB) 1999 recommendation maintained the receiver operating characteristic (ROC) curve cutoff for the diagnosis of AMI [28]. After the 2000 redefinition of MI due to the suboptimal analytical precision at the 99th percentile URL concentration for many of the cardiac troponin assays, the 10% CV concentration cutoff was proposed (i.e., the cardiac troponin concentration with an imprecision of $\leq 10\%$) [29]. This cutoff was selected to minimize the number of false-positive results due to analytical noise and enabled the differentiation of patients with stable versus unstable CAD. Improvements in analytical sensitivity and precision have occurred with the release of the next-generation cardiac troponin assays, whereby the 10% CV concentration level is at or near the 99th percentile URL of a healthy population, with the vast majority of the contemporary, sensitive assays achieving a CV $<20\%$ at the 99th percentile URL concentration [19].

Optimum cutoff concentrations

The use of sensitive cardiac troponin assays with the lower cutoff concentration (i.e., 99th percentile concentration versus the higher WHO ROC curve concentration) enables an earlier detection of AMI and identifies more individuals who are at risk in the short term (i.e., 30 days) for adverse cardiac events. MacRae et al. [30] demonstrated using a sensitive cardiac troponin assay at the 99th percentile with a change in criteria that measurement at least 3 hours after presentation gave equivalent AMI prevalence to samples collected at least 6 hours apart (Fig. 30.6). Melanson et al. [31] also showed that earlier detection of myocardial injury was possible with a sensitive cardiac troponin assay, and estimated that a sensitive cardiac troponin assay will detect injury at least 30 minutes earlier compared with the previous generation cardiac troponin assay. In the risk stratification study of TACTICS (Treat Angina with Aggrastat and Determine Cost of Therapy with an Invasive or Conservative Strategy)–Thrombolysis in Myocardial Infarction (TIMI) 18, Morrow et al. [32] identified 10% more patients who were at risk of future CVD relative to the 10% CV cutoff with standard cardiac troponin assays, without degradation in the hazard ratio. This was followed by another study by Morrow and colleagues, this time using a sensitive contemporary cardiac troponin I assay and the 99th percentile URL. Their findings in more than 4000 ACS patients were that patients with cardiac troponin concentrations between the 99th percentile concentration and the

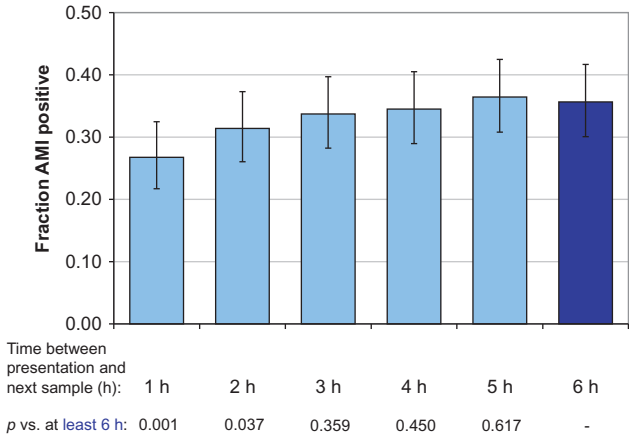


FIGURE 30.6 Effect of the minimum time interval between specimens on acute myocardial infarction positivity. Acute myocardial infarction positivity with progressively shorter minimal intervals between specimens compared with the sample taken at least 6 h after presentation. *From A.R. MacRae, P.A. Kavsak, V. Lustig, R. Bhargava, R. Vandersluis, G.E. Palomaki, et al., Assessing the requirement for the 6-hour interval between specimens in the American Heart Association Classification of Myocardial Infarction in Epidemiology and Clinical Research Studies. Clin. Chem. 52 (2006) 812–18 with permission.*

older, higher cardiac troponin cutoff used with the previous generation of assays (i.e., low concentrations above the 99th but below old cutoffs) were at significantly higher risk of death/MI at 30 days (5.0% vs. 2.0%, $P = .001$) and death at 12 months (6.4% vs. 2.4%, $P = .005$) than were patients with cardiac troponin concentrations below the 99th percentile URL [33]. These findings are consistent with other health outcome-based studies and guidelines, demonstrating the benefit of using the 99th percentile URL in patients with ACS symptoms [9,16]. However, even low, detectable, cardiac troponin concentrations below the 99th percentile may provide long-term (i.e., >1 year) risk stratification information for both men and women with ACS symptoms [34] (Fig. 30.7). Future cardiac troponin assays (i.e., the hs-cTn assays) will have even lower limits of detection, coupled with better precision at these low concentrations, which will result in lower and different 99th percentile URLs [19,35].

Lowering the decision limit for a clinical laboratory test may come at the expense of clinical specificity. While cardiac troponin is released following injury to the heart, it is not a specific marker of AMI (which is a clinical diagnosis). There are other conditions and diseases that cause myocardial damage and release of cardiac troponin. Some of these conditions include HF, renal failure, myocarditis, pulmonary embolus, sepsis, and use of drugs that are toxic to the heart (e.g., trastuzumab and doxorubicin) [36,37]. Opponents of the sensitive and hs-cTn assays will argue that lowering the cutoff concentration will result in a higher incidence of abnormal cardiac troponin concentrations in patients without AMI. It is important to emphasize that the diagnosis of AMI in the presence of low-concentration increases in cardiac troponin must be accompanied by other evidence of myocardial ischemia.

In an effort to reduce emergency department (ED) overcrowding, some physicians admitted patients to a coronary care unit on the basis of single positive cardiac troponin test result. This is not in compliance with the previous and current AMI guidelines that have consistently recommended serial testing to document a rise and fall in biomarkers. The use of a change in serial cardiac troponin results can be very helpful to determine if there is an acute cardiac injury (rising pattern), an ongoing release (stable pattern), or resolving damage (falling pattern). Patients with early onset ACS would be expected to have a rising pattern (or falling pattern if the onset of chest pain was distant), but there can be other causes of myocardial injury besides AMI. A patient with an increased cardiac troponin that is unchanging might have a disease that is associated with continual cellular damage. The biological variability is important in properly interpreting serial cardiac troponin results and has been reported for the hs-cTn assays. Based on the analytical and intraindividual coefficients of variances and log transformations, the reference change value (RCV) for cardiac troponin is a rise of about 50% and a fall of about 30%, as reported in one study [38]. Serial values in the short term, which are outside these limits, may imply acute cardiac injury. For sensitive cardiac troponin assays, some studies have derived relative percent change or absolute changes; however, these metrics are dependent on the assay used and the time between the measurements [39,40]. Also it should be stated that the derivation of the 99th percentile URL is specific for each cardiac troponin assay, dependent on how the "healthy population" was selected, and the number of individuals included to obtain a reliable estimate [41–44].

Other biomarkers of acute myocardial infarction

The success and incorporation of cardiac troponin into international guidelines and recommendations for ACS have put into question the need and value of existing biomarkers such as CK-MB, myoglobin, and LD, and other biomarkers of injury such as heart-type fatty acid binding protein (hFABP). While CK-MB has been in use as a cardiac biomarker for over 40 years, its incremental value relative to cardiac troponin has been questioned in the published ESC 4th Universal Definition of Myocardial Infarction [8]. CK and CK-MB with a molecular mass of 84 kDa has a monophasic release pattern from the cytoplasm within the first 12 hours after onset and remains increased for 2–3 days (Fig. 30.5A). While there has

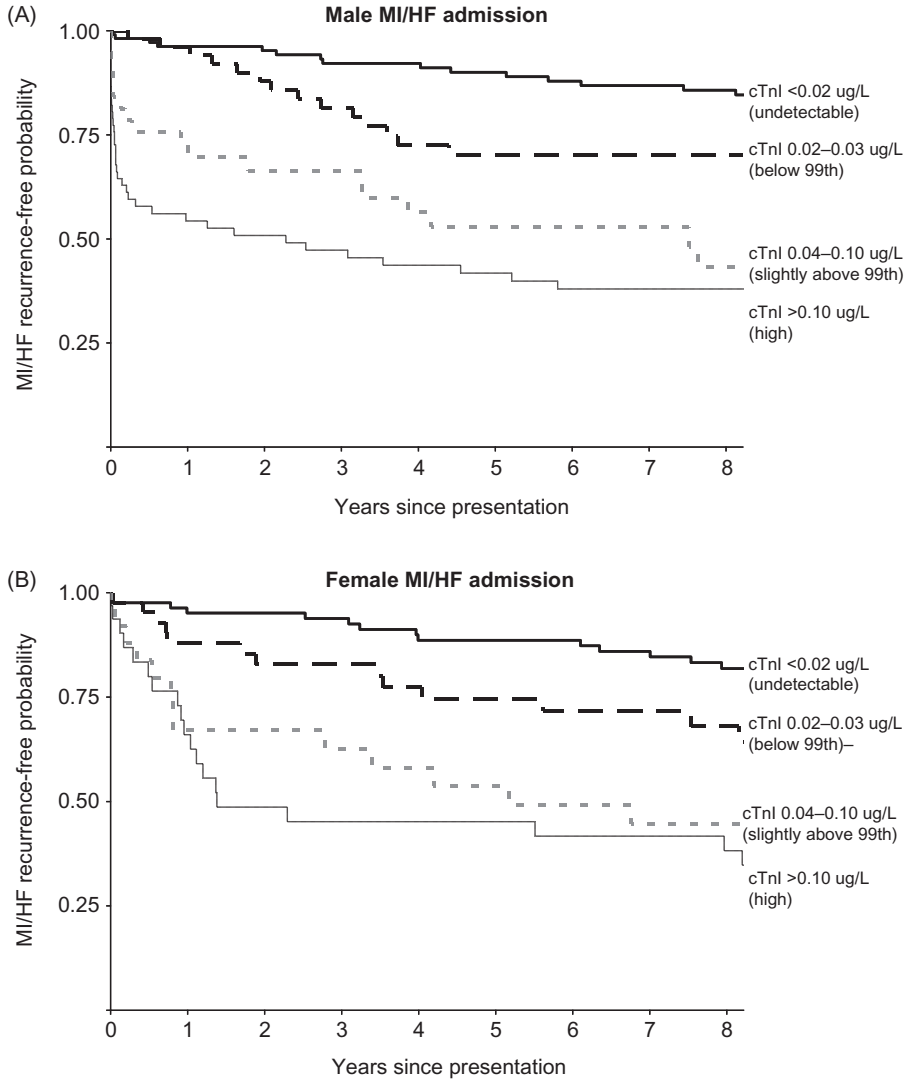


FIGURE 30.7 Kaplan–Meier curves for each sex for myocardial infarction/heart failure readmission endpoint from a population of patients who presented to the emergency department with symptoms suggestive of cardiac ischemia. Kaplan–Meier curves for myocardial infarction/heart failure readmission endpoint among the (A) 265 men and (B) 183 women (event-free survival in women: 82% for not detectable, 64% for low, 45% for intermediate, and 35% for high; event-free survival in men: 85%, 71%, 43%, and 38%, respectively; $P < .001$). *From P.A. Kavsak, A.M. Newman, V. Lustig, A.R. MacRae, G.E. Palomaki, D.T. Ko, et al., Long-term health outcomes associated with detectable troponin I concentrations. Clin. Chem. 53 (2007) 220–27 with permission.*

been some suggestion that CK appears earlier than cardiac troponin, this is clearly not the case when sensitive cardiac troponin cutoff concentrations are used, that is, at the 99th percentile [45]. Cardiac troponin is released earlier and remains increased for longer than CK-MB. Because of its more rapid return to baseline levels, some have suggested that while cardiac troponin is still increased, CK-MB could be used to detect the presence of reinfarction. However, Apple and Murakami [46] have shown that a second cardiac troponin peak can be observed in these patients, even if there is no return to baseline levels. Others have suggested monitoring serial CK-MB measurements for a determination of infarct size. This logic is flawed, as the objective for treating AMI patients is to revascularize as soon as an AMI is diagnosed to salvage jeopardized myocardium, that is, not to wait for a rise and fall to document the extent of injury. Successful reperfusion of a previously occluded coronary artery (as in the case of Type 1 MIs) invalidates the estimates of infarct sizing based on biomarker measurements.

Automated immunoassays for serum myoglobin have been available for the past 20 years as an early marker of AMI. This low-molecular-weight protein (~17 kDa) appears in blood within 6 hours after onset, reliably before CK-MB. The major limitation of myoglobin is the lack of cardiac specificity. Patients with renal failure and skeletal muscle injury will exhibit high myoglobin concentrations. While popular for a few years, the interest in myoglobin has waned as clinical studies have indicated inferior clinical performance as compared with sensitive cardiac troponin assays [45,47,48]. hFABP (encoded by the FABP3 gene) is another small protein that was originally shown to be released from myocardium that underwent ischemia and reperfusion with findings from a meta-analysis, suggesting that the combination of hFABP and sensitive cardiac troponin improved the sensitivity for

AMI as compared with cardiac troponin alone [49]. However, the data on the additive value of hFABP or copeptin (another potential biomarker used for early decision-making in the ED) when combined with hs-cTn assays for early AMI diagnosis are less favorable [50–52]. This is not to say there is no utility of the above biomarkers, as there may be other indications for measuring these biomarkers in different cardiac populations at risk of additional adverse events. One such example is in the setting of cardiac surgery where perioperative measurement of hFABP for predicating acute kidney injury appears to be superior to cardiac troponin in both the pediatric and adult cardiac surgery populations [53–55]. In addition, interferences that may be present in a patient's sample, such as heterophile or autoimmune antibodies to one particular cardiac troponin assay, may necessitate testing with a different cardiac troponin assay or different cardiac biomarkers in these situations or in rare diseases (an example being patients with neuromuscular myopathies and measurable troponin T) [26,27,56,57].

High-sensitivity cardiac troponin assays

The next generation of hs-cTn assays represents another major milestone for cardiac testing. Earlier versions including the sensitive cardiac troponin assays when measured in healthy, nonacutely ill, populations yielded undetectable cardiac troponin concentrations in the majority of individuals. With the advent of hs-cTn assays, now the majority of the healthy, nonacute, populations have measureable concentrations of cardiac troponin, with common consensus indicating that at least 50% of the population must have a measurable concentration with the cardiac troponin assay for it to be deemed a high-sensitivity assay [19,43,58]. However, the most recent laboratory practice recommendation from the Academy of the American Association for Clinical Chemistry (AACC) (previously named NACB) and the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), published in 2018, proposed that assays unable to detect cardiac troponin at concentrations at or above the limit of detection in at least 50% of healthy men and women be labeled as contemporary assays [44]. This concept is extrapolated from the fact that differences in concentrations between healthy men and women are evident when using hs-cTn assays and in addition to the sex-specific 99th percentiles being reported; those high-sensitivity assays should be able to detect measurable concentrations in both the male and female populations.

Concurrent with this, increase in the analytical sensitivity (approximately 10-fold more analytically sensitive than the sensitive/contemporary cardiac troponin assays) is an improved precision such that concentrations near and below the 99th percentile can be measured with a 10% CV. It should be noted that with this increased analytical performance that there needs to be an increased vigilance to specimen requirements, monitoring the assay and reporting results. For example, matrix differences (i.e., serum versus heparin plasma versus EDTA plasma) may exist (Fig. 30.8) and common interferences and preanalytical handling may affect measurements in one sample type or assay and not others [59–64], as concentrations below the 99th percentile can be measured with studies showing clinical utility in this range. Further, laboratories will also need additional quality control (QC) material and/or procedures in place to monitor hs-cTn assays in the "normal range" to prevent inappropriate drifts in signal at the low end due to instrument performance, reagent lot stability, or new lots of reagent [65,66] (Fig. 30.9). The need for a "normal" concentration QC material when offering hs-cTn testing has been endorsed by the AACC/IFCC expert group, in that their recommendation 1 for hs-cTn assays indicates a QC material above

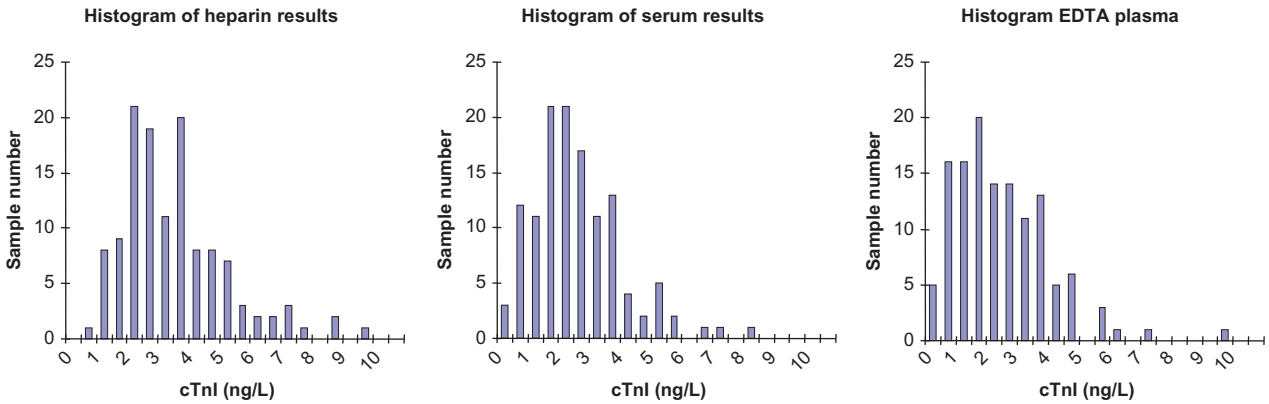


FIGURE 30.8 Distribution of cardiac troponin concentrations in 125 healthy individuals in different matrices with a high-sensitivity cardiac troponin assay. *From P.A. Kavsak, A.R. MacRae, M.J. Yerna, A.S. Jaffe, Analytic and clinical utility of a next-generation, highly sensitive cardiac troponin I assay for early detection of myocardial injury. Clin. Chem. 55 (2009) 573–77 with permission.*

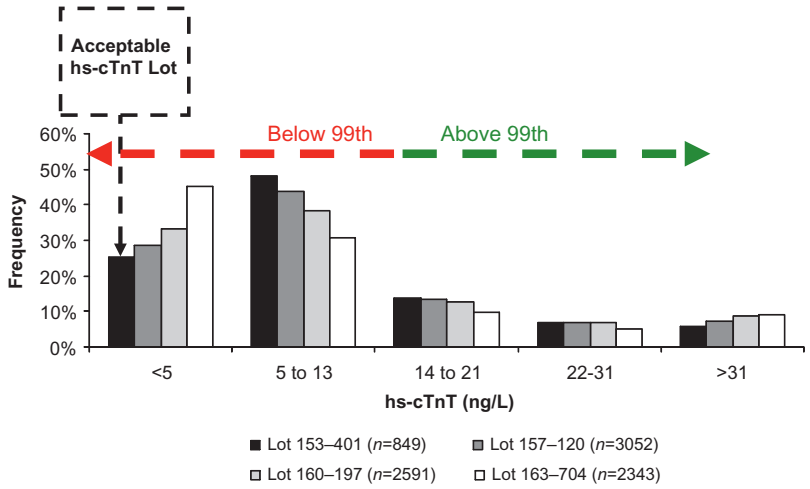


FIGURE 30.9 Distribution of high-sensitivity cardiac troponin T results with one acceptable lot (153401) versus different lots of reagents (March 2010 to September 2012). The percentages of high-sensitivity cardiac troponin T concentrations below the limit of detection (<5 ng/L) were higher in lots 157120 (29%; $P = .032$), 160197 (33%; $P < .001$), and 163704 (45%; $P < .001$), compared with the acceptable lot 153401 (25%). *From P.A. Kavsak, S.A. Hill, M.J. McQueen, P.J. Devereaux, Implications of adjustment of high-sensitivity cardiac troponin T assay. Clin. Chem. 59 (2013) 574–76 with permission.*

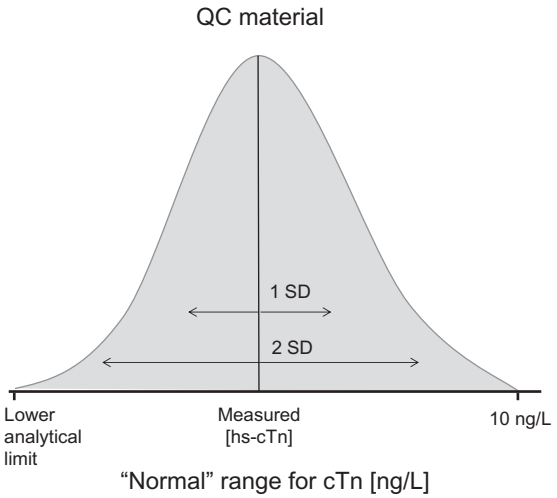


FIGURE 30.10 An illustration demonstrating the utility of measuring a quality control material in the "normal" concentration range for high-sensitivity cardiac troponin assays (i.e., above the lower limit of reporting and below 10 ng/L). The SD value should be ≤0.8 ng/L for high-sensitivity cardiac troponin measurements below 10 ng/L [68].

the limit of detection and below the lowest sex-specific 99th percentile [44]. This is in addition to a QC material near the highest 99th percentile and another QC material at concentrations multiple-fold higher than the 99th percentile to challenge the upper analytical range [44]. Importantly, having a QC material below 10 ng/L (a concentration range that is measurable with high-sensitivity assays), precision goals and total error criteria that have been developed with hs-cTn assays may be utilized [44,67–69] (Fig. 30.10).

Other recommendations by the AACC/IFCC expert group include the validation of the lower limit of reporting, which may be the limit of detection or limit of quantification [44]. Different regulatory bodies may impose or limit concentrations that can be reported with hs-cTn assays [18,44]. Accordingly, standardization and agreement of testing at the low-concentration end may be quite variable [70]. Consensus has also emerged that clinically hs-cTn results should be reported in whole units and in ng/L [8,19,43,44]. This will help avoid confusion among users and may even help in the clinical implementation of hs-cTn assays [44,71]. Another recommendation by the expert group that may help laboratories in offering the most sensitive and precise cardiac troponin assay is the change in turnaround time, where cardiac troponin results should be reported within 60 minutes or less of when a sample is received in the laboratory [44]. This represents a change from previous recommendations that indicate 60 minutes from blood collection, with the rationale being that laboratories can appropriately track, monitor, and control turnaround times when samples are received in the laboratory. In all, the AACC/IFCC expert group has listed 10 recommendations to help laboratories when offering cardiac troponin testing, with a focus on hs-cTn assays [44].

Clinically, hs-cTn assays have advantages over contemporary and sensitive cardiac troponin assays in the early decision-making process in the ED and in the nonacute setting. Specifically, studies have demonstrated that undetectable or very low cardiac troponin concentrations in patients with possible ACS at ED presentation may be ruled out for AMI or earlier serial measurements before the 3–6 hour time frame may be useful for early ED

disposition (i.e., 1- or 2-hour algorithms), though there are important laboratory quality issues and population related aspects that need to be taken into consideration in these approaches [72–77]. Using the published ESC guidelines yields high negative and positive predictive values for AMI, yet the sensitivity of these algorithms may not be sufficient for all physicians and for all patient populations [77–81]. Moreover, despite population reference intervals clearly indicating differences between the 99th percentiles for men and women, a large stepped-wedge, cluster-randomized controlled trial assessing sex-specific 99th percentile cutoffs with one hs-cTn assay found that more patients were identified with myocardial injury than type 1 MI and this did not lead to a reduction in subsequent cardiac events [82]. However, these same authors reported that an hs-cTn assay can identify low-risk patients [82]. This is in agreement with observations obtained in stable, nonacute CVD populations, where hs-cTn assays are able to identify patients at low and high risks for long-term adverse cardiovascular outcomes and death [83–87]. Importantly, data are emerging that treatment decisions based on hs-cTn in the nonacute setting may be beneficial to patient care [87–89]. Accordingly, the use of hs-cTn assays is becoming mainstream, with several hs-cTn assays already approved for clinical use [18,64,69], with more in development [90]. This, therefore, diminishes the need for the contemporary assays and obviates the advantages of other nontroponin biomarkers in the early ED decision-making process as shorter ordering protocols and diagnostic and prognostic evidences have been demonstrated with hs-cTn assays [91].

Pathophysiology of heart failure

HF is a clinical syndrome with patients often presenting with fatigue, shortness of breath, exercise intolerance, and fluid retention; however, as not all patients present with volume overload, the term HF is preferred over congestive heart failure (CHF) [92]. HF is a chronic CVD that occurs when the myocardium is unable to supply oxygen-enriched blood to peripheral tissues to meet metabolic demands and results from impairment of ventricular filling or ejection due to disorders of the myocardium, endocardium, pericardium, heart valves, and great vessels or from other metabolic abnormalities [92]. Risk factors for HF include CAD (and ACS/MI), hypertension, diabetes mellitus, and metabolic syndrome; with the left ventricular ejection fraction (LVEF) an important variable for classification [HF with preserved EF (HFpEF) and HF with reduced EF (HFrEF)] [92]. HFrEF (LVEF $\leq$ 40%), also known as systolic HF, has been most commonly studied with efficacious therapies established in clinical studies, while HFpEF (LVEF $\geq$ 50%) also known as diastolic HF is diagnostically more challenging with limited therapies [92]. The ESC Guidelines have also classified LVEF ranges from 40% to 49% as HF with mid-range EF, as these patients may have different underlying pathophysiology [93]. Typically, HFpEF versus HFrEF patients are older, more commonly female, with greater hypertension, obesity, anemia, atrial fibrillation, and lower natriuretic peptide concentrations [94,95]. In summary, HF is a clinical diagnosis based on history and physical examination, though laboratory tests may be helpful in making this diagnosis.

In HF, the pathophysiological result of reduced systemic blood flow leads to the activation of the renin–aldosterone axis, which causes vasoconstriction, sodium and water retention, and hypertension (Fig. 30.11, right side). In a time of crisis, this axis is necessary to maintain blood pressure and blood flow to vital organs including the brain. Unfortunately, chronic overstimulation of this axis that occurs in HF causes cardiac enlargement and remodeling (Fig. 30.12). A shift in the shape from an elliptical to a spherical geometry makes the heart less efficient in pumping blood out to the body, thereby causing a deficit between blood supply and demand. As a

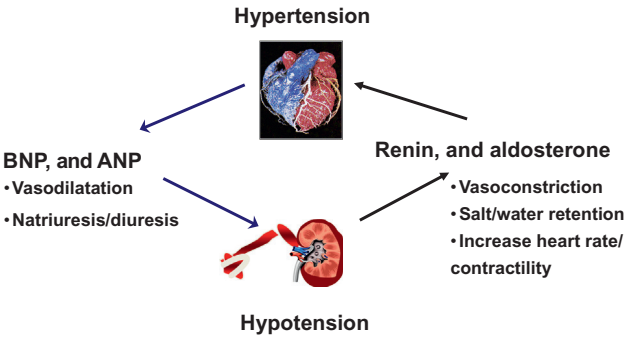


FIGURE 30.11 The hormonal balance between renin–aldosterone and atrial and b-type natriuretic peptide. *ANP*, Atrial natriuretic peptide; *BNP*, B-type natriuretic peptide.

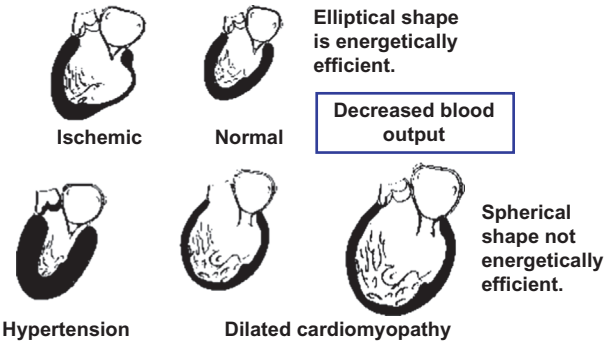


FIGURE 30.12 The chronic overstimulation of the renin–aldosterone axis dilated cardiomyopathy leads to cardiac enlargement and remodeling. The shift from an elliptical to a spherical geometry causes an inefficiency in cardiac output leading to heart failure.

compensatory mechanism, electrolyte imbalance and the stretching of the myocardium caused by cardiomegaly stimulate the release of atrial (ANP) and BNP and are used as biomarkers for this condition. These hormones counteract the renin–aldosterone axis by vasodilation, sodium and water loss, and hypotension (Fig. 30.11, left side). The natriuretic peptides, in various molecular forms, have become important indicators for the diagnosis, staging, and monitoring of HF.

B-type natriuretic peptide and N-terminal-proBNP

Biochemistry

The biochemistry of the natriuretic peptides is complex. ANP has the highest tissue content in the atrium and is stored in cytoplasmic granules and can be quickly released following the acute and transient volume changes. BNP comes from the ventricle and must be upregulated at the genomic level. As HF is most associated with the impaired left ventricular function and sustained volume overload, BNP is more widely used as a clinical biomarker. BNP originates from an inactive precursor protein pre-proBNP (134 amino acids) with the signal peptide (26 amino acids) cleaved to yield the prohormone proBNP (108 amino acids). Upon stimulation, $proBNP_{108}$, a 12-kDa peptide, is cleaved stoichiometrically by the proteases corin and/or furan to BNP_{1-32} (active hormone, ~4 kDa) and the NT-$proBNP_{1-76}$ (inactive peptide, ~8 kDa, see Fig. 30.13) [96]. The half-life of BNP is roughly 20 minutes, and for NT-proBNP, it is between 1 and 2 hours, with NT-proBNP also having higher in vitro stability than BNP. Normal ranges for BNP and NT-proBNP increase from each decade of life, with women having higher levels than men. Obese subjects have lower BNP and NT-proBNP concentrations than lean individuals. In addition to these forms, proBNP itself is released into circulation and is believed to be less biologically active/inactive than mature BNP. The proBNP and the corresponding amino acids contained within NT-proBNP undergo glycosylation. The clinical significance of this posttranslational modification is unknown, though a study has indicated that blood samples treated with deglycosylation enzymes yielded NT-proBNP concentrations that were higher (versus nontreated) when measured with a clinical assay [97]. In addition, intracellular glycosylation on threonine 71 may prevent cleavage of proBNP between the amino acids 76 and 77 by furin and corin [96,98].

Commercial assays are available for BNP and NT-proBNP. Like the different cardiac troponin assays, there are minor differences in the clinical utility of BNP and NT-proBNP. A clinical laboratory will use one test or the other depending on the availability of the commercial testing platform currently in use in that laboratory. It is not fully known at this point whether the clinical utility of proBNP is identical to or provides incremental value to BNP and NT-proBNP.

Clinical utility of B-type natriuretic peptide and N-terminal-proBNP in heart failure

Diagnosis

BNP or NT-proBNP are routinely used for the diagnosis of HF among ED and hospitalized patients. There have been two landmark studies that have examined the clinical utility of BNP for patients who present to an ED with acute dyspnea and shortness of breath. In the multicenter BNP trial, a BNP cutoff of 100 ng/L (pg/mL) produced a clinical sensitivity for HF of 90% and specificity of 76%

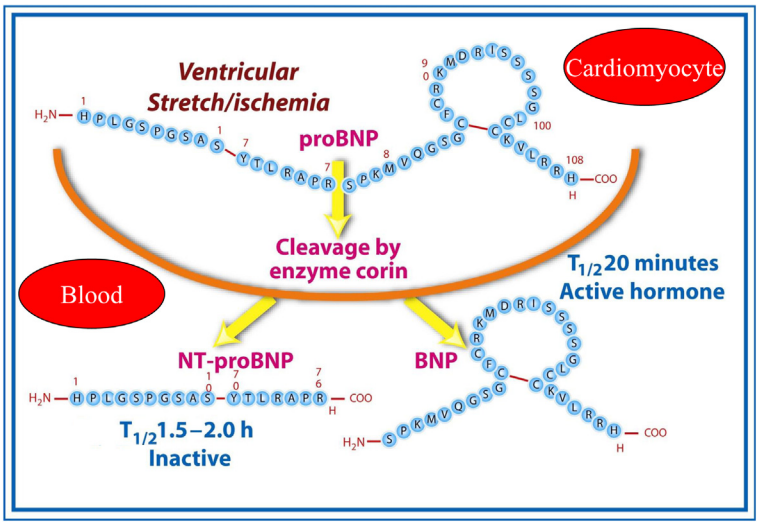


FIGURE 30.13 The processing by corin and subsequent release of proBNP into the circulation as B-type natriuretic peptide (active hormone) and N-terminal-proBNP (inactive peptide).

[99]. From the proBNP Investigation of Dyspnea in the Emergency Department (PRIDE) study, similar clinical performance was observed for NT-proBNP using a cutoff of 450 ng/L for those <50 years, 900 ng/L for those ≥50 years, with an NT-proBNP concentration <300 ng/L optimal for ruling-out acute HF, with a negative predictive value of 99% [100]. Both studies identified a "gray zone" area, as elevations may be caused by non-HF etiologies such as renal failure, hypertension, and pulmonary disease. Thus different concentration ranges have been recommended: (1) for BNP <100 ng/L, HF is unlikely; (2) for BNP concentrations between 100 and 500 ng/L, HF is possible but other diagnoses need to be considered; and (3) with BNP >500 ng/L, HF is very likely. The same concepts apply for the NT-proBNP assay using different cutpoints based on age [95]. To this end, the ICON: Re-evaluation of Acute Diagnostic Cut-Offs in the Emergency Department study has validated the originally proposed age-specific NT-proBNP cutoffs [i.e., age-specific rule-in cutoffs of 450 ng/L (<50 years), 900 ng/L (50−75 years), and 1800 ng/L (>75 years) for the diagnosis of acute HF, and <300 ng/L as the rule-out cutoff to exclude diagnosis of acute HF] in a contemporary cohort of dyspneic patients presenting to EDs in North America [101].

Clinical societies have also endorsed natriuretic peptide testing for the diagnosis of HF in both the acute and ambulatory settings (typically class I recommendation with level of evidence of A) [92,95], with systematic reviews supporting natriuretic peptide testing for ruling-out HF [102]. The 2016 ESC guidelines for the diagnosis and treatment of acute and chronic HF list natriuretic peptide cutoffs in the nonacute setting (BNP <35 ng/L and NT-proBNP <125 ng/L) and in the acute setting (BNP <100 ng/L and NT-proBNP <300 ng/L) [93]. Whereas the 2017 ACC/AHA/HFSA Focused Update of the 2013 ACCF/AHA Guideline for the Management of Heart Failure has not listed specific cutoffs, the update has provided the following Class 1, Level A recommendations for natriuretic peptide testing: (1) in patients presenting with dyspnea, measurement of natriuretic peptide biomarkers is useful to support a diagnosis or exclusion of HF; (2) measurement of BNP or NT-proBNP is useful for establishing prognosis or disease severity in chronic HF; and (3) measurement of baseline levels of natriuretic peptide biomarkers and/or cardiac troponin on admission to the hospital is useful to establish a prognosis in acutely decompensated HF [103]. However, this update has provided natriuretic cutoff information in the following class I recommendation with respect to treatment: in patients with chronic symptomatic HFrEF NYHA class II or III who tolerate an ACE inhibitor or angiotensin receptor blocker (ARB), replacement by an angiotensin receptor-neprilysin inhibitors (ARNI) is recommended to reduce further morbidity and mortality (103, with ARNI effects on BNP versus NT-proBNP concentrations discussed below under the *Monitoring and Management* subsection).

Staging

Assays for BNP/NT-proBNP can also be useful for staging the severity of disease using criteria such as the New York Heart Association Classification (Fig. 30.14) and use of the 6-minute walk test. While a laboratory test is not needed to determine that a patient is in Stage IV (symptoms of shortness of breath at rest), this test is particularly useful for the detection of Stage I HF (asymptomatic patients). This has led some investigators to suggest using the natriuretic peptides for screening asymptomatic individuals. However, Wang et al. [104] reviewed existing studies and concluded that mass screening with echocardiography or the natriuretic peptides was not cost-effective. It may be possible to improve the utility by screening subjects who are at the highest risk for HF. For example, the St Vincent's screening to prevent heart failure randomized trial assessing BNP-based screening and collaborative care on patients with cardiovascular risk factors (i.e., patients at risk for HF) demonstrated benefit with the study approach likely being cost-effective as well [105,106]. Despite some randomized controlled trials in this area, the 2017 ACC/AHA/HFSA Focused Update of the 2013 ACCF/AHA Guidelines have provided a Class IIa (moderate) Level B-R (moderate from one or more RCTs) recommendation for natriuretic peptide testing for the prevention of HF with further studies needed to assess cost-effectiveness, quality of life, and mortality [103].

Monitoring and management

A novel application of the natriuretic peptides is the use of these tests to support therapeutics in HF. There have been numerous studies examining the potential to use

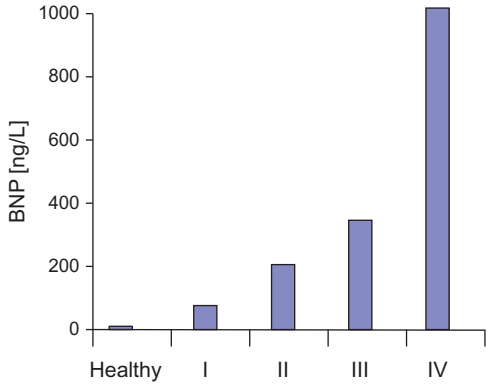


FIGURE 30.14 Correlation of B-type natriuretic peptide concentrations with New York Heart Association classification of heart failure.

serial natriuretic peptide testing to monitor the success of drug therapy such as with diuretics and inhibitors to angiotensin-converting enzyme and angiotensin II. Interpretation of serial results requires knowledge of the biological variation and limits for RCV. Roughly, a 75%−100% increase and a 40%−50% decrease are necessary to determine statistical change in BNP/NT-proBNP concentrations. In patients admitted for decompensated HF, levels of BNP/NT-proBNP can decline by greater than these serial limits. For stable HF outpatients, it is more difficult to produce decreases in natriuretic peptide levels to below these limits. Nevertheless, several prospective trials have attempted to improve HF therapy by increasing anti-HF doses to achieve a target BNP or NT-proBNP concentration. In the STARS-BNP trial (supported by an unrestricted grant from Biosite Inc), patients in the BNP group who had a goal of decreasing BNP to less than 100 ng/L had significantly fewer combined endpoints (unplanned hospital endpoints, death, and hospital stay for HF) than in the group managed without BNP measurements [107]. The NT-proBNP-assisted treatment to lessen serial cardiac readmissions and death study using NT-proBNP showed similar results but only in a subset of younger patients [108], and the findings from the trial of intensified vs. standard medical therapy in elderly patients with congestive heart failure study suggest that NT-proBNP-guided therapy may be more beneficial to HFrEF patients than to HFpEF patients [109]. Therefore from these trials, systematic reviews, and meta-analyses, natriuretic peptide testing-guided therapy showed no benefit in older patients (>75 years) and in HFpEP patients [94,95,102]. Until more studies are reported unequivocally demonstrating clinical benefit to targeting drug therapy to specific BNP or NT-proBNP concentrations, caution is warranted; however, the findings of the Guiding Evidence Based Therapy Using Biomarker Intensified Treatment (GUIDE-IT) trial—Clinical Trial registration: NCT01685840 should be able to assess the effects of a natriuretic peptide testing strategy in high-risk HFrEF patients [110]. Unfortunately, the GUIDE-IT trial in high-risk patients with HFrEF found that a strategy of NT-proBNP-guided therapy was not more effective than a usual care strategy in improving outcomes, with a potential contributing variable to the null findings being the intensive treatment/follow-up in the control arm [111].

Moreover, different beneficial treatments may result in differences between BNP and NT-proBNP concentrations while patients are on these treatments. Specifically, the ARNI [ARB and neprilysin inhibitor, example drug brand name Entresto (sacubitril (neprilysin inhibitor)/valsartan (ARB)] drug class has been shown to be more effective in preventing death, cardiovascular death, and clinical progression of surviving HF patients than angiotensin-converting enzyme inhibition, with the premise that blocking neprilysin (an enzyme that breaks down bradykinin, adrenomedullin, and natriuretic peptides) enhanced the effects of these peptides in HF patients [112]. To this end, HF patients treated with an ARNI may expect an increase in BNP concentrations (action of the drug), while NT-proBNP concentrations may drop representing the beneficial effects of the drug [112]. These important effects of ARNIs on BNP and NT-proBNP concentrations need to be clearly communicated to the health-care team and patients in order to avoid confusion in this setting. To this end, the 2017 ACC expert consensus decision pathway for the optimization of HF treatment has indicated that clinicians need to interpret natriuretic peptides in the context of guideline-directed management and therapy and in the case of ARNI therapy; perhaps, it is more prudent to check only NT-proBNP concentrations [113].

Clinical utility of B-type natriuretic peptide and N-terminal-proBNP in acute coronary syndrome

While BNP and NT-proBNP are considered markers of chronic left ventricular dysfunction, there is considerable evidence that these peptides are released during the initial stages of myocardial ischemia as well. Measurement of the natriuretic peptides in the setting of acute chest pain offers prognostic information for adverse cardiac events, particularly when used in conjunction with other biomarkers. The combination of BNP with cardiac troponin and C-reactive protein (CRP), a marker of acute and chronic inflammation, appears particularly powerful [114]. Fig. 30.15 shows the role of these biomarkers in three cases of MI, which differ in location and etiology. The cardiac troponin concentration is increased in all cases as per the Universal Definition of MI. Fig. 30.15A shows an AMI caused by vasospasm, such as with an acute intoxication with cocaine. As plaque instability and rupture is not the mechanism for AMI, there may not be an inflammatory role for this infarction and CRP may be normal. The natriuretic peptides are increased due to the effect that the myocardial damage has on contractile function. Fig. 30.15B shows an AMI due to plaque rupture occurring at the end of the right coronary artery, near the apex of the heart. Plaque rupture suggested some underlying etiology of inflammation and an increase in CRP. Since this region of the myocardium does not have a major contractile role in pumping out blood, this infarction may not cause a rise in BNP/NT-proBNP. Fig. 30.15C shows an infarct caused by an occlusion of the left anterior descending artery, a key vessel for delivery of blood to the left ventricle. Here, all three blood biomarkers are increased, as there is inflammation and damage to the left regional wall.

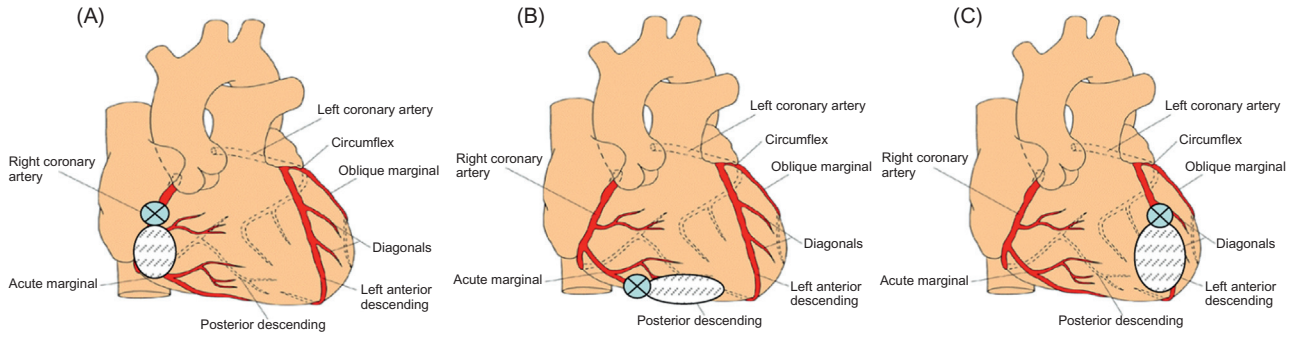


FIGURE 30.15 The combination of natriuretic peptides, cardiac troponin, and C-reactive protein as a multimarker tool for risk assessment in acute coronary disease. The underlying mechanism and location of the infarct could affect which markers are increased and the severity of the increase. See text for details. Site of occlusion. Area of myocardial necrosis. *LAD*, Left anterior descending coronary artery.

TABLE 30.3 Future needs of cardiac biomarkers in acute coronary syndrome and heart failure.

Acute coronary syndrome
- Early detection of myocardial infarction at ED presentation (soon after and ideally even before pain onset)
- Risk stratification for short-term major adverse events (absence of cardiac troponin elevation)
- Differentiation between Type 1 MI and Type 2 MI

Heart failure
- Adds incremental value to natriuretic peptides when results are in the "gray zone" (e.g., renal failure and obesity).
- Improved prediction of cardiovascular morbidity and mortality
- Monitoring and selection of appropriate therapy in both HFrEF and HFpEF patients

ED, Emergency department; *HFpEF*, heart failure with preserved ejection fraction; *HFrEF*, heart failure with reduced ejection fraction; *MI*, myocardial infarction.

Multimarker analysis has potential to play an important role as the clinical cardiac problems and conditions become more complicated (e.g., Type 2 MI and HFpEP). There are additional cardiac conditions that cause elevated natriuretic peptide concentrations besides HF and ACS, with heart muscle disease, valvular heart disease, pericardial disease, atrial fibrillation, myocarditis, toxicity, cardioversion, and cardiac surgery being potential causes [103]. In addition, on the noncardiac side, anemia, renal failure, pulmonary hypertension and disease (i.e., severe pneumonia and obstructive sleep apnea), critical illness, bacteria sepsis, severe burns, and advancing age may also result in elevations [103].

Clinical need for future biomarkers of acute cardiovascular diseases

Table 30.3 lists the future needs for biomarkers within the areas of ACS and HF [115]. In the ED, there is a need to have a cardiac biomarker with high clinical sensitivity and specificity for ACS at the time of patient presentation. For both ACS and HF, better tools are needed to predict morbidity and mortality, particularly if they can be linked to novel therapeutics.

Early acute coronary syndrome diagnosis

Fig. 30.16 shows the classic sequence of biomarker release for cardiac troponin and the natriuretic peptides, biomarkers of myocardial cell injury/necrosis, and ventricular overload/remodeling. None of these biomarkers are reliably increased at or immediately after the onset of symptoms/chest pain, with early rule-out strategies using hs-cTn assays in early presenters also being suboptimal [73,77]. Therefore a laboratory test that can detect the presence of myocardial ischemia, which precedes irreversible damage, would be valuable. Table 30.4 summarizes some of the biomarkers that have been studied for myocardial ischemia (free fatty acids, glycogen phosphorylase BB isoenzyme, ischemia modified albumin, and

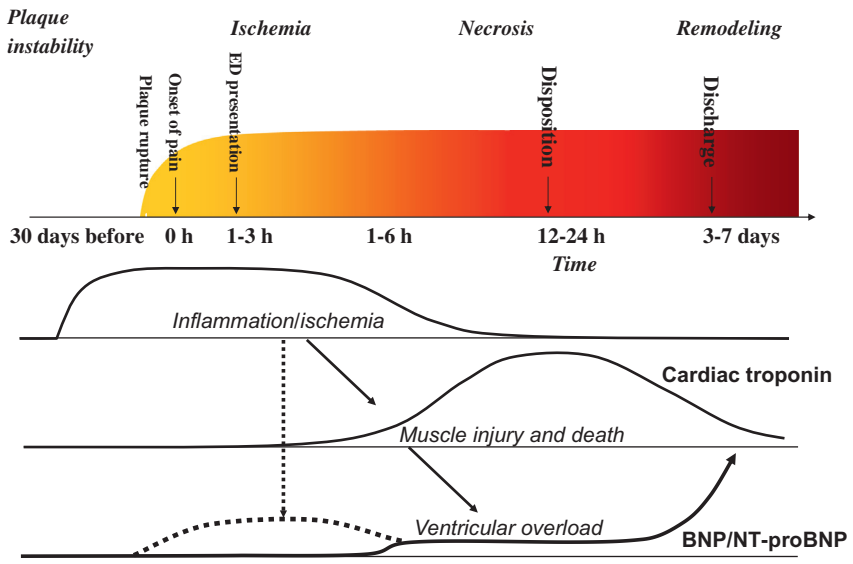


FIGURE 30.16 Timing of current and novel biomarkers in acute coronary syndrome. Cardiac troponin is increased early (in hours) from onset of pain (plaque rupture). B-type natriuretic peptide/N-terminal-proBNP is increased days later with myocardial remodeling. There is also an ischemic release of B-type natriuretic peptide/N-terminal-proBNP. Other blood-based biomarkers are needed to detect the onset of myocardial ischemia and immediately before the onset of Type I myocardial infarction (plaque instability).

TABLE 30.4 Other potential biomarkers of acute coronary syndrome and heart failure.

Marker	Description	Diagnostic role
Acute coronary syndrome		
Free fatty acid	Unbound serum fatty acid	Myocardial ischemia
Glycogen phosphorylase BB	Enzyme in glycogenolysis	Myocardial ischemia
Ischemia modified albumin	N-terminal modification by ischemia	Myocardial ischemia
Plasma choline	Product of phospholipase activation	Myocardial ischemia
Whole blood choline	Product of phospholipase activation	Plaque rupture
Soluble CD40 ligand	Circulating ligand of the CD40 receptor	Plaque rupture
Placental-like growth factor	Vascular endothelial growth factor	Plaque instability
PAPP-A	Metalloproteinase	Plaque instability
Myeloperoxidase	Prooxidant enzyme of leukocytes	Plaque instability
MCP-1	Recruits monocytes into the vasculature	Plaque instability
Copeptin	C-terminal portion of provasopressin	Early rule-out
Heart failure		
ST2	Isoform of IL-1 receptor/fibrosis	Predicator of cardiac death
proAdrenomedullin	Precursor to vasodilative natriuretic peptide	Predictor of cardiac death
proBNP	Precursor to BNP and NT-proBNP	Indicator of impaired natriuretic peptide processing
Galectin-3	β-galactosidase binding lectin/fibrosis	Predictor of cardiac death
GDF-15	TGF-β member/biochemical strain	Predictor of cardiac death

BNP, B-type natriuretic peptide; *GDF*, growth differentiation factor; *MCP-1*, Monocyte chemoattractant protein-1; *NT*, N-terminal; *PAPP-A*, Pregnancy-associated plasma protein A.

plasma choline), and monocyte (whole blood choline) and platelet activation (CD40 ligand) that are indicative of plaque rupture [115]. While these biomarkers are released during the early phases of AMI, the clinical specificity toward cardiac ischemia and thrombosis is not high. Many of the biochemical events that occur within the heart are in common with other organs such as the brain, lungs, intestinal tract, and peripheral tissues. Since cardiac troponin, especially hs-cTn, can be used to rule-out AMI in a few hours after ED presentation, many ED physicians are willing to wait for the more definitive test result rather than to base an early discharge decision on the use of a noncardiac biomarker. This decision is particularly important in the US, as wrongful discharge of a patient with ACS is a frequent cause of malpractice litigation from the ED today. However, large meta-analyses evaluating either hs-cTnT or hs-cTnI alone to rule-out MI via the presentation blood sample using 5 ng/L as the cutoff have yielded very high negative predictive values; however, the sensitivity at this cutoff may not be acceptable to all

physicians [116,117]. Other approaches coupling hs-cTn with other laboratory tests or clinical scores may further improve the sensitivity and rule-out [118,119].

Improved risk stratification for adverse cardiac events

After a patient has been ruled out for ACS in the ED, the next objective is to determine the risk of that patient for developing an adverse cardiac event within the next 30 days, a meaningful time frame for ED physicians. Cardiac troponin is used today to risk stratify patients at high risk. Unfortunately, up to 5% of patients with a negative (below cutoff) cardiac troponin concentration will also suffer AMI or cardiac death, and additional risk stratification tools are needed. A major area of research has been in detecting the presence of plaque vulnerability as in the case of Type 1 MI. Fig. 30.16 shows that such a marker would predict plaque rupture, ischemia, necrosis, and remodeling. Table 30.4 lists some of the tests that have been examined in clinical studies. They are focused on the pathophysiologic events that make plaques vulnerable to rupture, including inflammation (e.g., myeloperoxidase), fibrous cap degradation (pregnancy-associated plasma protein A), angiogenesis (placental-like growth factor, by providing more vascular access to harmful factors), and leukocyte chemotaxis (monocyte chemoattractant protein-1, stimulus to recruit harmful leukocytes to the site of damage). A combination of biomarkers that detect plaque instability, plaque rupture, ischemia, necrosis, and left ventricular dysfunction may be very useful for future diagnoses and risk stratification in ACS. More work will be needed to validate such an approach and to provide commercial tools for the clinical laboratory, such as integrated testing platforms, and test reporting algorithms that are cost-effective. These multivariate decision-making tools must be separately cleared by the US FDA prior to their clinical use. Most importantly, there must be directed therapeutics driven by clinical laboratory tests that can be given to high-risk patients to reduce their short-term CVD risk. This may require a partnership with the pharmaceutical industry to create "companion" diagnostic tests.

Other heart failure biomarkers

While BNP and NT/proBNP are routinely used in the diagnosis and evaluation of HF, their limitations have led some investigators to examine novel biomarkers. ST2 (also known as IL1RL1, DER4, T1 and FIT-1) was originally discovered by mechanically stressing myocytes in cell culture to simulate the progression of HF and use of proteomics to discover the biomarkers that are up- or downregulated. Subsequently, ST2 has been determined to be an isoform of the interleukin-1 receptor, which is secreted into the circulation and binds interleukin-33 [120]. ST2 is less influenced by obesity than BNP or NT-proBNP. Adrenomedullin (ADM) is a natriuretic and vasodilative hormone produced by the heart [121]. The mid-region (MR)-pro-adrenomedullin (proADM) is a more stable precursor peptide to ADM. In the PRIDE and biomarkers in acute heart failure studies, increased concentrations of ST2 and MR-proADM were strong and independent predictors of mortality among HF patients. These results were additive to the value of BNP and NT-proBNP. MR-proADM has a lower biological variation than BNP/NT-proBNP, making this test potentially more useful for therapeutic monitoring [122]. ProBNP is the precursor protein to BNP and NT-proBNP and is released in HF [123]. Preliminary data suggest that this marker may provide insights to patients with impaired natriuretic processing, that is, are unable to produce BNP in severe HF. While it is unlikely that these analytes will replace BNP and/or NT-proBNP in routine practice, the addition of these biomarkers in a multimarker approach may provide incremental clinical value, as they represent different pathophysiological processes. In fact, in addition to ST2, another fibrosis biomarker that has also obtained US FDA approval is galectin-3 [124], with both biomarkers being mentioned in the 2013 ACCF/AHA Guideline for the Management of Heart Failure [92]. Another potential biomarker in this area is growth differentiation factor-15 (GDF-15), a member of the transforming growth factor β cytokine superfamily, which is expressed in the models of MI, pressure overload, and HF [124,125]

Laboratory analysis for cardiac biomarkers

Assays for cardiac troponin and the natriuretic peptides are based on two-site sandwich immunoassays, which are available on automated immunoassay testing platforms and point-of-care-testing-device (though there is a single epitope BNP assay approved in China; see IFCC Committee on Clinical Applications of Cardiac Bio-Markers (C-CB) website that lists tables for cardiac troponin and the natriuretic peptides providing information on assay design and other analytical parameters: http://www.ifcc.org/ifcc-education-division/emd-committees/task-force-on-clinical-applications-of-cardiac-bio-markers-tf-cb/) [90]. Most commercial cardiac troponin assays have undergone a series of improvements to reduce interferences and imprecision and to increase analytical sensitivity, which has also been the case for the hs-cTn assays [126]. The most widely used assays for cardiac troponin T are through Roche Diagnostics, while there are multiple manufacturers of cardiac troponin I assays [19], for an up to date listing see IFCC C-CB

website [90]. Test results differ between and even within manufactures, as manufactures may produce more than one cardiac troponin assay [127]. There is an ongoing effort to standardize/harmonize cardiac troponin testing [128]; however, the differences in antibodies used and methods used in commercial cardiac troponin tests may prohibit complete standardization of results. In fact, there are even differences between the hs-cTn tests as compared with the non-hs-cTn tests from the same manufacturer, as often the assay needs to undergo further optimization to become a high-sensitivity assay (i.e., increase sample volume, antibody reconfiguration, buffers, etc.), so results from one assay to another assay may be divergent [59,129,130]. Point-of-care-testing platforms are also available for testing whole blood and enable a more rapid turnaround time for reporting results and may reduce the length of stay for ED patients with chest pain [131]. Point-of-care-testing devices for cardiac biomarkers are generally not as sensitive or precise as current generation central laboratory equipment, resulting in a reduced ability to detect individuals at risk of adverse events [132,133]. However, advancements are being made for point-of-care hs-cTn testing [134].

Central laboratory and point-of-care-testing devices are also available for BNP and NT-proBNP. Unlike cardiac troponin, low concentrations of these peptides in blood have little clinical significance; therefore, there is not the same need to develop high-sensitivity assays though the ESC guidelines have recommended a BNP cutoff of 35 ng/L in the nonacute setting, indicating that precision is still important at the lower end. There is no standard reference material for BNP or NT-proBNP. Harmonization between assays is not as poor as for cardiac troponin, especially for NT-proBNP. Many of the NT-proBNP assays are licensed from Roche Diagnostics who requires the use of the same calibrators and antibodies and results are similar. Although there are minor differences in BNP assays, most manufacturers have established the same cutoff concentration of 100 ng/L to be used in the acute setting. However, important preanalytical and sample handling issues must be adhered to for BNP measurement, as EDTA plasma is the only suitable matrix for BNP measurement, with BNP more likely to undergo degradation than NT-proBNP [135]. An IFCC laboratory educational report published in 2019 on natriuretic peptide testing with a focus on HF has listed 10 recommendations on the analytical and clinical issues affecting interpretation of the results, with a new proposed target CV $<10\%$ being stated to improve the analytical performance [136].

Conclusions

Cardiac biomarkers continue to play a critical role in the management of patients with CVD. The clinical sensitivity and specificity for cardiac troponin for AMI is among the highest for any clinical laboratory test. The value of natriuretic peptide testing continues to grow from a role in HF diagnosis to prognosis to treatment management decisions. As illustrated in this chapter, there is continued research into new and other biomarkers for heart disease, some cardiac in nature and others not. The biomarkers that can assist in the diagnosis and management of patients with stroke remain elusive with ongoing effort looking into biomarkers that reflect endothelial damage, coagulation/fibrinolysis, and inflammation/growth factors with interest in multimarker panels as well [137]. This will be a much bigger challenge than the myocardial injury due to ischemic processes, as biomarkers evident of brain injury may need to cross the blood−brain barrier, and more analytically, sensitivity assays may be required to detect these products. There is hope here, however, as one only needs to look at the timeline of cardiac biomarkers [4], and in particular cardiac troponin where in 2016 (the previous version of this chapter in the Third Edition) there were no hs-cTn assays approved by the FDA for use in the US. Fast forward, to 2019 (at the time of this chapter revision), there are now many different manufacturers with hs-cTn assay approved for use by the FDA in the US [18,63,90,127,138].

References

[1] D. Mozaffarian, E.J. Benjamin, A.S. Go, D.K. Arnett, M.J. Blaha, M. Cushman, et al., Heart disease and stroke statistics--2015 update: a report from the American Heart Association, Circulation 131 (2015) e29−e322.

[2] GBD 2013 Mortality and Causes of Death Collaborators, Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990−2013: a systematic analysis for the Global Burden of Disease Study 2013, Lancet 385 (2015) 117−171.

[3] J.S. LaDue, F. Wroblewski, A. Karmen, Serum glutamic oxaloacetic transaminase activity in human acute transmural myocardial infarction, Science 120 (1954) 497−499.

[4] P.O. Collinson, L. Garrison, R.H. Christenson, Cardiac biomarkers - a short biography, Clin. Biochem. 48 (2015) 197−200.

[5] M. Naghavi, P. Libby, E. Falk, S.W. Casscells, S. Litovsky, J. Rumberger, et al., From vulnerable plaque to vunerable patient. A call for new definitions and risk assessment strategies: part I, Circulation 108 (2003) 1664−1672.

[6] P. Libby, Current concepts of the pathogenesis of the acute coronary syndrome, Circulation 104 (2001) 365−372.

[7] Nomenclature and Criteria for Diagnosis of Ischemic Heart Disease Report of the Joint International Society and Federation of Cardiology/World Health Organization task force on standardization of clinical nomenclature. Circulation 1979;59:607-609.

[8] K. Thygesen, J.S. Alpert, A.S. Jaffe, B.R. Chaitman, J.J. Bax, D.A. Morrow, et al., Fourth universal definition of myocardial infarction (2018), Eur. Heart J. 40 (2019) 237−269.

[9] K. Thygesen, J.S. Alpert, A.S. Jaffe, M.L. Simoons, B.R. Chaitman, H.D. White, et al., Third universal definition of myocardial infarction, Circulation 126 (2012) 2020–2035.
[10] A.S. Jaffe, Third universal definition of myocardial infarction, Clin. Biochem. 46 (2013) 1–4.
[11] Vascular Events in Noncardiac Surgery Patients Cohort Evaluation (VISION) Study Investigators, Association between postoperative troponin levels and 30-day mortality among patients undergoing noncardiac surgery, JAMA 307 (2012) 2295–2304.
[12] P.M. Ridker, Clinician's Guide to reducing inflammation to reduce atherothrombotic risk: JACC review topic of the week, J. Am. Coll. Cardiol. 72 (2018) 3320–3331.
[13] A.H.B. Wu, Biochemical markers of cardiac damage: from traditional enzymes to cardiac-specific proteins, Scan J. Clin. Lab. Invest. 59 (suppl) (1999) 74–82.
[14] Joint European Society of Cardiology/American College of Cardiology Committee, Myocardial infarction redefined—a consensus document of the joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction, J. Am. Coll. Cardiol. 36 (2000) 959–969.
[15] K. Thygesen, J.S. Alpert, H.D. White, the Joint ESC/ACCF/AHA/WHF Task Force for the Redefinition of Myocardial Infarction, Universal definition of myocardial infarction, Circulation 116 (2007) 2634–2653.
[16] P.A. Kavsak, A.R. MacRae, G.E. Palomaki, A.M. Newman, D.T. Ko, V. Lustig, et al., Health outcomes categorized by current and previous definitions of acute myocardial infarction in an unselected cohort of troponin-naïve emergency department patients, Clin. Chem. 52 (2006) 2028–2035.
[17] N.L. Mills, A.M. Churchhouse, K.K. Lee, A. Anand, D. Gamble, A.S. Shah, et al., Implementation of a sensitive troponin I assay and risk of recurrent myocardial infarction and death in patients with suspected acute coronary syndrome, JAMA 305 (2011) 1210–1216.
[18] P.A. Kavsak, J.E. Andruchow, A.D. McRae, A. Worster, Profile of Roche's Elecsys Troponin T Gen 5 STAT blood test (a high-sensitivity cardiac troponin assay) for diagnosing myocardial infarction in the emergency department, Expert. Rev. Mol. Diagn. 18 (2018) 481–489.
[19] F.S. Apple, P.O. Collinson, IFCC Task Force on Clinical Applications of Cardiac Biomarkers, Analytical characteristics of high-sensitivity cardiac troponin assays, Clin. Chem. 58 (2012) 54–61.
[20] P. Hwang, A. Mingels, P.A. Kavsak, High-sensitivity cardiac troponin testing during and after ACS: complexed or not? Clin. Biochem. (2019). pii: S0009-9120(19)30762-3.
[21] I.A. Katrukha, A.E. Kogan, A.V. Vylegzhanina, M.V. Serebryakova, E.V. Koshkina, A.V. Bereznikova, et al., Thrombin-mediated degradation of human cardiac troponin T, Clin. Chem. 63 (2017) 1094–1100.
[22] I.A. Katrukha, A.E. Kogan, A.V. Vylegzhanina, A.V. Kharitonov, N.N. Tamm, V.L. Filatov, et al., Full-size cardiac troponin I and its proteolytic fragments in blood of patients with acute myocardial infarction: antibody selection for assay development, Clin. Chem. 64 (2018) 1104–1112.
[23] S. Zahran, V.P. Figueiredo, M.M. Graham, R. Schulz, P.M. Hwang, Proteolytic digestion of serum cardiac troponin I as marker of ischemic severity, J. Appl. Lab. Med. 3 (2018) 450–455.
[24] P.E. Hickman, J.M. Potter, C. Aroney, G. Koerbin, E. Southcott, A.H. Wu, et al., Cardiac troponin may be released by ischemia alone, without necrosis, Clin. Chim. Acta 411 (2010) 318–323.
[25] F. Sedaghat-Hamedani, E. Kayvanpour, L. Frankenstein, D. Mereles, A. Amr, S. Buss, et al., Biomarker changes after strenuous exercise can mimic pulmonary embolism and cardiac injury–a metaanalysis of 45 studies, Clin. Chem. 61 (2015) 1246–1255.
[26] A.S. Jaffe, V.C. Vasile, M. Milone, A.K. Saenger, K.N. Olson, F. S. Apple, Diseased skeletal muscle: a noncardiac source of increased circulating concentrations of cardiac troponin T, J. Am. Coll. Cardiol. 58 (2011) 1819–1824.
[27] P.A. Kavsak, C. Roy, P. Malinowski, et al., Macrocomplexes and discordant high-sensitivity cardiac troponin concentrations, Ann. Clin. Biochem. 55 (2018) 500–504.
[28] A.H. Wu, F.S. Apple, W.B. Gibler, R.L. Jesse, M.M. Warshaw, R. Valdes Jr., National Academy of Clinical Biochemistry Standards of Laboratory Practice: recommendations for the use of cardiac markers in coronary artery diseases, Clin. Chem. 45 (1999) 1104–1121.
[29] F.S. Apple, A.H. Wu, A.S. Jaffe, European Society of Cardiology and American College of Cardiology guidelines for redefinition of myocardial infarction: how to use existing assays clinically and for clinical trials, Am. Heart J. 144 (2002) 981–986.
[30] A.R. MacRae, P.A. Kavsak, V. Lustig, R. Bhargava, R. Vandersluis, G.E. Palomaki, et al., Assessing the requirement for the 6-hour interval between specimens in the American Heart Association Classification of Myocardial Infarction in Epidemiology and Clinical Research Studies, Clin. Chem. 52 (2006) 812–818.
[31] S.E. Melanson, D.A. Morrow, P. Jarolim, Earlier detection of myocardial injury in a preliminary evaluation using a new troponin I assay with improved sensitivity, Am. J. Clin. Pathol. 128 (2007) 282–286.
[32] D.A. Morrow, C.P. Cannon, N. Rifai, M.J. Frey, R. Vicari, N. Lakkis, et al., Ability of minor elevations of troponins I and T to predict benefit from an early invasive strategy in patients with unstable angina and non-ST elevation myocardial infarction: results from a randomized trial, JAMA 286 (2001) 2405–2412.
[33] M. Bonaca, B. Scirica, M. Sabatine, A. Dalby, J. Spinar, S.A. Murphy, et al., Prospective evaluation of the prognostic implications of improved assay performance with a sensitive assay for cardiac troponin I, J. Am. Coll. Cardiol. 55 (2010) 2118–2124.
[34] P.A. Kavsak, A.M. Newman, V. Lustig, A.R. MacRae, G.E. Palomaki, D.T. Ko, et al., Long-term health outcomes associated with detectable troponin I concentrations, Clin. Chem. 53 (2007) 220–227.
[35] F.S. Apple, R. Ler, M.M. Murakami, Determination of 19 cardiac troponin I and T assay 99th percentile values from a common presumably healthy population, Clin. Chem. 58 (2012) 1574–1581.
[36] A.H.B. Wu, F.S. Apple, A.S. Jaffe, R.L. Jesse, D.A. Morrow, K. Newby, et al., National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines: use of cardiac troponin and the natriuretic peptides for etiologies other than acute coronary syndromes and heart failure, Clin. Chem. 53 (2007) 2086–2096.
[37] W.E. Kelley, J.L. Januzzi, R.H. Christenson, Increases of cardiac troponin in conditions other than acute coronary syndrome and heart failure, Clin. Chem. 55 (2009) 2098–2112.

[38] A.H.B. Wu, A. Lu, J. Todd, J. Moecks, F. Wians, Short- and long-term biological variation for cardiac troponin I using a high sensitivity assay: implications for clinical practice, Clin. Chem. 55 (2009) 52–58.
[39] P.A. Kavsak, D.T. Ko, X. Wang, A.R. MacRae, A.S. Jaffe, 2007 universal myocardial infarction definition change criteria for risk stratification by use of a high-sensitivity cardiac troponin I assay, Clin. Chem. 56 (2010) 487–489.
[40] T. Reichlin, A. Irfan, R. Twerenbold, M. Reiter, W. Hochholzer, H. Burkhalter, et al., Utility of absolute and relative changes in cardiac troponin concentrations in the early diagnosis of acute myocardial infarction, Circulation 124 (2011) 136–145.
[41] Y. Sandoval, F.S. Apple, The global need to define normality: the 99th percentile value of cardiac troponin, Clin. Chem. 60 (2014) 455–462.
[42] P.A. Kavsak, A. Rezanpour, Y. Chen, K. Adeli, Assessment of the 99th or 97.5th percentile for cardiac troponin I in a healthy pediatric cohort, Clin. Chem. 60 (2014) 1574–1576.
[43] F.S. Apple, A.S. Jaffe, P. Collinson, M. Mockel, J. Ordonez-Llanos, B. Lindahl, et al., IFCC educational materials on selected analytical and clinical applications of high sensitivity cardiac troponin assays, Clin. Biochem. 48 (2015) 201–203.
[44] A.H.B. Wu, R.H. Christenson, D.N. Greene, A.S. Jaffe, P.A. Kavsak, J. Ordonez-Llanos, et al., Clinical laboratory practice recommendations for the use of cardiac troponin in acute coronary syndrome: expert opinion from the Academy of the American Association for Clinical Chemistry and the Task Force on clinical applications of cardiac bio-markers of the International Federation of Clinical Chemistry and Laboratory Medicine, Clin. Chem. 64 (2018) 645–655.
[45] T. Keller, T. Zeller, D. Peetz, S. Tzikas, A. Roth, E. Czyz, et al., Sensitive troponin I assay in early diagnosis of acute myocardial infarction, N. Engl. J. Med. 361 (9) (2009) 868–877.
[46] F.S. Apple, M.M. Murakami, Cardiac troponin and creatine kinase MB monitoring during in-hospital myocardial reinfarction, Clin. Chem. 51 (2005) 460–463.
[47] K.M. Eggers, J. Oldgren, A. Nordenskjold, B. Lindahl, Diagnostic value of serial measurement of cardiac markers in patients with chest pain: limited value of adding myoglobin to troponin I for exclusion of myocardial infarction, Am. Heart J. 148 (2004) 574–581.
[48] P.A. Kavsak, A.R. MacRae, A.M. Newman, V. Lustig, G.E. Palomaki, D.T. Ko, et al., Effects of contemporary troponin assay sensitivity on the utility of the early markers myoglobin and CKMB isoforms in evaluating patients with possible acute myocardial infarction, Clin. Chim. Acta 380 (2007) 213–216.
[49] G. Lippi, C. Mattiuzzi, G. Cervellin, Critical review and meta-analysis on the combination of heart-type fatty acid binding protein (H-FABP) and troponin for early diagnosis of acute myocardial infarction, Clin. Biochem. 46 (2013) 26–30.
[50] K.M. Eggers, P. Venge, B. Lindahl, High-sensitive cardiac troponin T outperforms novel diagnostic biomarkers in patients with acute chest pain, Clin. Chim. Acta 413 (2012) 1135–1140.
[51] C.R. Shortt, A. Worster, S.A. Hill, P.A. Kavsak, Comparison of hs-cTnI, hs-cTnT, hFABP and GPBB for identifying early adverse cardiac events in patients presenting within six hours of chest pain-onset, Clin. Chim. Acta 419 (2013) 39–41.
[52] P. Hillinger, R. Twerenbold, C. Jaeger, K. Wildi, T. Reichlin, M. Rubini Gimenez, et al., Optimizing early rule-out strategies for acute myocardial infarction: utility of 1-hour copeptin, Clin. Chem. 61 (2015) 1466–1474.
[53] E.M. Bucholz, R.P. Whitlock, M. Zappitelli, P. Devarajan, J. Eikelboom, A.X. Garg, et al., Cardiac biomarkers and acute kidney injury after cardiac surgery, Pediatrics 135 (2015) e945–e956.
[54] J.A. Schaub, A.X. Garg, S.G. Coca, J.M. Testani, M.G. Shlipak, J. Eikelboom, et al., Perioperative heart-type fatty acid binding protein is associated with acute kidney injury after cardiac surgery, Kidney Int. 88 (2015) 576–583.
[55] P.A. Kavsak, R. Whitlock, H. Thiessen-Philbrook, C.R. Parikh, Perioperative heart-type fatty acid binding protein concentration cutoffs for the identification of severe acute kidney injury in patients undergoing cardiac surgery, Clin. Chem. Lab. Med. 57 (2018) e8–e10.
[56] J.Y. Park, A.S. Jaffe, Troponin autoantibodies: from assay interferent to mediator of cardiotoxicity, Clin. Chem. 63 (2017) 30–32.
[57] D.S. Herman, P.A. Kavsak, D.N. Greene, Variability and error in cardiac troponin testing: an ACLPS critical review, Am. J. Clin. Pathol. 148 (2017) 281–295.
[58] P.A. Kavsak, L.C. Allen, F.S. Apple, R.A. Booth, P.C. Chan, E. Delvin, et al., Cardiac troponin testing in the acute care setting: ordering, reporting, and high sensitivity assays--an update from the Canadian society of clinical chemists (CSCC), Clin. Biochem. 44 (2011) 1273–1277.
[59] P.A. Kavsak, A.R. MacRae, M.J. Yerna, A.S. Jaffe, Analytic and clinical utility of a next-generation, highly sensitive cardiac troponin I assay for early detection of myocardial injury, Clin. Chem. 55 (2009) 573–577.
[60] J. Simons, L. Beach, L. Clark, P.A. Kavsak, Matrix and bilirubin interference for high-sensitivity cardiac troponin I, Clin. Chim. Acta 442 (2015) 49–51.
[61] P.A. Kavsak, N. Caruso, J. Beattie, L. Clark, Centrifugation--an important pre-analytical factor for the Abbott Architect high-sensitivity cardiac troponin I assay, Clin. Chim. Acta 436 (2014) 273–275.
[62] P.A. Kavsak, P. Malinowski, C. Roy, L. Clark, S. Lamers, Assessing matrix, interferences and comparability between the Abbott Diagnostics and the Beckman Coulter high-sensitivity cardiac troponin I assays, Clin. Chem. Lab. Med. 56 (2018) 1176–1181.
[63] P.A. Kavsak, A. Worster, S.A. Hill, A.S. Jaffe, Evaluation of the Siemens ADVIA Centaur high-sensitivity cardiac troponin I assay in serum, Clin. Chim. Acta 487 (2018) 216–221.
[64] A.K. Saenger, A.S. Jaffe, R. Body, P.O. Collinson, P.A. Kavsak, C.S.P. Lam, et al., Cardiac troponin and natriuretic peptide analytical interferences from hemolysis and biotin: educational aids from the IFCC Committee on Cardiac Biomarkers (IFCC C-CB), Clin. Chem. Lab. Med. 57 (2019) 633–640.
[65] P.A. Kavsak, S.A. Hill, M.J. McQueen, P.J. Devereaux, Implications of adjustment of high-sensitivity cardiac troponin T assay, Clin. Chem. 59 (2013) 574–576.
[66] P.A. Kavsak, A. Worster, R. Oliver, L. Clark, D. Parry, E. Randell, et al., Variability between reagent lots for high-sensitivity cardiac troponin I may affect performance of early rule out strategies, Can. J. Cardiol. 34 (2018) 209.e5–209.e6.
[67] P.A. Kavsak, A.C. Don-Wauchope, S.A. Hill, A. Worster, Acceptable analytical variation may exceed high-sensitivity cardiac troponin I cutoffs in early rule-out and rule-in acute myocardial infarction algorithms, Clin. Chem. 62 (2016) 887–889.
[68] P.A. Kavsak, A.S. Jaffe, D.N. Greene, R.H. Christenson, F.S. Apple, A.H.B. Wu, Total analytic error for low cardiac troponin concentrations (≤10 ng/L) by use of a high-sensitivity cardiac troponin assay, Clin. Chem. 63 (2017) 1043–1045.

[69] P.A. Kavsak, Should detectable cardiac troponin concentrations in a healthy population be the only criterion for classifying high-sensitivity cardiac troponin assays? Clin. Biochem. 56 (2018) 1−3.
[70] P.A. Kavsak, External quality assessment testing near the limit of detection for high-sensitivity cardiac troponin assays, Clin. Chem. 64 (2018) 1402−1404.
[71] H. Kufaishi, A. Pardhan, W. Krizmanich, A. Worster, S. Hill, P.A. Kavsak, Adopting 'ng/L' as the units for high-sensitivity cardiac troponin assays and commitment by the entire health-care team could be the key for adopting recommendations, Ann. Clin. Biochem. 53 (2016) 516−517.
[72] R. Body, G. Burrows, S. Carley, L. Cullen, M. Than, A.S. Jaffe, et al., High-sensitivity cardiac troponin T concentrations below the limit of detection to exclude acute myocardial infarction: a prospective evaluation, Clin. Chem. 61 (2015) 983−989.
[73] A.S. Shah, A. Anand, Y. Sandoval, K.K. Lee, S.W. Smith, P.D. Adamson, et al., High-sensitivity cardiac troponin I at presentation in patients with suspected acute coronary syndrome: a cohort study, Lancet 386 (2015) 2481−2488.
[74] P.A. Kavsak, A. Worster, Dichotomizing high-sensitivity cardiac troponin T results and important analytical considerations, J. Am. Coll. Cardiol. 59 (17) (2012) 1570.
[75] T. Reichlin, C. Schindler, B. Drexler, R. Twerenbold, M. Reiter, C. Zellweger, et al., One-hour rule-out and rule-in of acute myocardial infarction using high-sensitivity cardiac troponin T, Arch. Intern. Med. 172 (2012) 1211−1218.
[76] P.A. Kavsak, High-five for high-sensitivity cardiac troponin T: depends on the precision and analytical platform, JAMA Intern. Med. 173 (2013) 477.
[77] M. Roffi, C. Patrono, J.P. Collet, C. Mueller, M. Valgimigli, F. Andreotti, et al., 2015 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: Task Force for the Management of Acute Coronary Syndromes in patients presenting without persistent ST-segment elevation of the European Society of Cardiology (ESC), Eur. Heart J. 37 (2016) 267−315.
[78] J.W. Pickering, J.H. Greenslade, L. Cullen, D. Flaws, W. Parsonage, P. George, et al., Validation of presentation and 3h high-sensitivity troponin to rule-in and rule-out acute myocardial infarction, Heart 102 (2016) 1270−1278.
[79] J.W. Pickering, J.H. Greenslade, L. Cullen, D. Flaws, W. Parsonage, S. Aldous, et al., Assessment of the European Society of Cardiology 0-hour/1-hour algorithm to rule-out and rule-in acute myocardial infarction, Circulation 134 (2016) 1532−1541.
[80] R. Twerenbold, J.T. Neumann, N.A. Sörensen, F. Ojeda, M. Karakas, J. Boeddinghaus, et al., Prospective validation of the 0/1-h algorithm for early diagnosis of myocardial infarction, J. Am. Coll. Cardiol. 72 (2018) 620−632.
[81] J.S. Alpert, A.S. Jaffe, 1-h high-sensitivity troponin rule-out and rule-in approach: strengths and potential weaknesses, J. Am. Coll. Cardiol. 72 (2018) 633−635.
[82] A.S.V. Shah, A. Anand, F.E. Strachan, A.V. Ferry, K.K. Lee, A.R. Chapman, et al., High-sensitivity troponin in the evaluation of patients with suspected acute coronary syndrome: a stepped-wedge, cluster-randomised controlled trial, Lancet 392 (2018) 919−928.
[83] T. Omland, J.A. de Lemos, M.S. Sabatine, C.A. Christophi, M.M. Rice, K.A. Jablonski, et al., A sensitive cardiac troponin T assay in stable coronary artery disease, N. Engl. J. Med. 361 (2009) 2538−2547.
[84] P.A. Kavsak, L. Xu, S. Yusuf, M.J. McQueen, High-sensitivity cardiac troponin I measurement for risk stratification in a stable high-risk population, Clin. Chem. 57 (2011) 1146−1153.
[85] T. Omland, M.A. Pfeffer, S.D. Solomon, J.A. de Lemos, H. Røsjø, J. Šaltytė Benth, et al., PEACE Investigators. Prognostic value of cardiac troponin I measured with a highly sensitive assay in patients with stable coronary artery disease, J. Am. Coll. Cardiol. 61 (2013) 1240−1249.
[86] A.L. Beatty, I.A. Ku, R.H. Christenson, C.R. DeFilippi, N.B. Schiller, M.A. Whooley, High-sensitivity cardiac troponin T levels and secondary events in outpatients with coronary heart disease from the Heart and Soul Study, JAMA Intern. Med. 173 (2013) 763−769.
[87] S. Blankenberg, V. Salomaa, N. Makarova, F. Ojeda, P. Wild, K. J. Lackner, et al., Troponin I and cardiovascular risk prediction in the general population: the BiomarCaRE consortium, Eur. Heart J. 37 (2016) 2428−2437.
[88] I. Ford, A.S. Shah, R. Zhang, D.A. McAllister, F.E. Strachan, M. Caslake, et al., High-sensitivity cardiac troponin, statin therapy, and risk of coronary heart disease, J. Am. Coll. Cardiol. 68 (2016) 2719−2728.
[89] P.J. Devereaux, The potential for troponin to inform prognosis in patients with stable coronary artery disease, Ann. Intern. Med. 169 (2018) 808−809.
[90] P.O. Collinson, A.K. Saenger, F.S. Apple, High sensitivity, contemporary and point-of-care cardiac troponin assays: educational aids developed by the IFCC Committee on Clinical Application of Cardiac Bio-Markers, Clin. Chem. Lab. Med. 57 (2019) 623−632.
[91] J.T. Neumann, R. Twerenbold, F. Ojeda, N.A. Sörensen, A.R. Chapman, A.S.V. Shah, et al., Application of high-sensitivity troponin in suspected myocardial infarction, N. Engl. J. Med. 380 (2019) 2529−2540.
[92] C.W. Yancy, M. Jessup, B. Bozkurt, J. Butler, D.E. Casey Jr, M. H. Drazner, et al., 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, J. Am. Coll. Cardiol. 62 (2013) e147−e239.
[93] P. Ponikowski, A.A. Voors, S.D. Anker, H. Bueno, J.G.F. Cleland, A.J.S. Coats, et al., 2016 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) developed with the special contribution of the Heart Failure Association (HFA) of the ESC, Eur. Heart J. 37 (2016) 2129−2200.
[94] E. Braunwald, Heart failure, JACC Heart Fail. 1 (2013) 1−20.
[95] G.W. Moe, J.A. Ezekowitz, E. O'Meara, S. Lepage, J.G. Howlett, S. Fremes, et al., The 2014 Canadian Cardiovascular Society Heart Failure Management Guidelines Focus Update: anemia, biomarkers, and recent therapeutic trial implications, Can. J. Cardiol. 31 (2015) 3−16.
[96] T.G. Yandle, A.M. Richards, B-type natriuretic peptide circulating forms: analytical and bioactivity issues, Clin. Chim. Acta 448 (2015) 195−205.
[97] H. Røsjø, M.B. Dahl, M. Jørgensen, R. Røysland, J. Brynildsen, A. Cataliotti, et al., Influence of glycosylation on diagnostic and prognostic accuracy of N-terminal pro-B-type natriuretic peptide in acute dyspnea: data from the Akershus Cardiac Examination 2 Study, Clin. Chem. 6 (2015) 1087−1097.

[98] V.C. Vasile, A.S. Jaffe, Natriuretic peptides and analytical barriers, Clin. Chem. 63 (2017) 50–58.
[99] A.S. Maisel, P. Krishnaswamy, R.M. Nowak, J. McCord, J.E. Hollander, P. Duc, et al., Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure, N. Engl. J. Med. 347 (2002) 161–167.
[100] J.L. Januzzi Jr., C.A. Camargo, S. Anwaruddin, A.L. Baggish, A. A. Chen, D.G. Karuser, et al., The N-terminal Pro-BNP investigation of dyspnea in the emergency department (PRIDE) study, Am. J. Cardiol. 95 (2005) 948–954.
[101] J.L. Januzzi Jr, A.A. Chen-Tournoux, R.H. Christenson, G. Doros, J.E. Hollander, P.D. Levy, et al., N-terminal pro-B-type natriuretic peptide in the emergency department: the ICON-RELOADED Study, J. Am. Coll. Cardiol. 71 (2018) 1191–1200.
[102] A.C. Don-Wauchope, R.S. McKelvie, Evidence based application of BNP/NT-proBNP testing in heart failure, Clin. Biochem. 48 (2015) 236–246.
[103] C.W. Yancy, M. Jessup, B. Bozkurt, J. Butler, D.E. Casey Jr, M. M. Colvin, 2017 ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Failure Society of America, J. Am. Coll. Cardiol. 70 (2017) 776–803.
[104] T.J. Wang, D. Levy, E.J. Benjamin, R.S. Vasan, The epidemiology of asymptomatic left ventricular systolic dysfunction: implications for screening, Ann. Intern. Med. 138 (2003) 907–916.
[105] M. Ledwidge, J. Gallagher, C. Conlon, E. Tallon, E. O'Connell, I. Dawkins, et al., Natriuretic peptide-based screening and collaborative care for heart failure: the STOP-HF randomized trial, JAMA 310 (2013) 66–74.
[106] M.T. Ledwidge, E. O'Connell, J. Gallagher, L. Tilson, S. James, V. Voon, et al., Cost-effectiveness of natriuretic peptide-based screening and collaborative care: a report from the STOP-HF (St Vincent's Screening TO Prevent Heart Failure) study, Eur. J. Heart Fail. 17 (2015) 672–679.
[107] P. Jourdain, G. Jondeau, F. Funck, P. Gueffet, A. Le Helloco, E. Donal, et al., Plasma brain natriuretic peptide-guided therapy to improve outcome in heart failure. The STARS-BNP Multicenter Study, J. Am. Coll. Cardiol. 49 (2007) 1733–1739.
[108] J.G. Lainchbury, R.W. Troughton, K.M. Strangman, C.M. Frampton, A. Pilbrow, T.G. Yandle, et al., N-terminal pro-B-type natriuretic peptide-guided treatment for chronic heart failure: results from the BATTLESCARRED (NT-proBNP-Assisted Treatment To Lessen Serial Cardiac Readmissions and Death) trial, J. Am. Coll. Cardiol. 55 (2009) 53–60.
[109] M.T. Maeder, P. Rickenbacher, H. Rickli, H. Abbühl, M. Gutmann, P. Erne, et al., N-terminal pro brain natriuretic peptide-guided management in patients with heart failure and preserved ejection fraction: findings from the Trial of Intensified versus standard medical therapy in elderly patients with congestive heart failure (TIME-CHF), Eur. J. Heart Fail. 15 (2013) 1148–1156.
[110] G.M. Felker, T. Ahmad, K.J. Anstrom, K.F. Adams, L.S. Cooper, J.A. Ezekowitz, et al., Rationale and design of the GUIDE-IT study: guiding evidence based therapy using biomarker intensified treatment in heart failure, JACC Heart Fail. 2 (2014) 457–465.
[111] G.M. Felker, K.J. Anstrom, K.F. Adams, J.A. Ezekowitz, M. Fiuzat, N. Houston-Miller, et al., Effect of natriuretic peptide-guided therapy on hospitalization or cardiovascular mortality in high-risk patients with heart failure and reduced ejection fraction: a randomized clinical trial, JAMA 318 (2017) 713–720.
[112] M. Packer, J.J. McMurray, A.S. Desai, J. Gong, M.P. Lefkowitz, A.R. Rizkala, et al., Angiotensin receptor neprilysin inhibition compared with enalapril on the risk of clinical progression in surviving patients with heart failure, Circulation 131 (2015) 54–61.
[113] C.W. Yancy, J.L. Januzzi Jr, L.A. Allen, J. Butler, L.L. Davis, G.C. Fonarow, et al., 2017 ACC expert consensus decision pathway for optimization of heart failure treatment: answers to 10 pivotal issues about heart failure with reduced ejection fraction: a report of the American College of Cardiology Task Force on expert consensus decision pathways, J. Am. Coll. Cardiol. 71 (2018) 201–230.
[114] M.S. Sabatine, D.A. Morrow, J.A. de Lemos, C.M. Gibson, S.A. Murphy, N. Rifai, et al., Multimarker approach to risk stratification in non-ST elevation acute coronary syndromes: simultaneous assessment of troponin I, C-reactive protein, and B-type natriuretic peptide, Circulation 105 (2002) 1760–1763.
[115] F.S. Apple, A.H.B. Wu, J. Mair, J. Ravkilde, M. Panteghini, J. Tate, et al., Future biomarkers for detection of ischemia and risk stratification in acute coronary syndrome, Clin. Chem. 51 (2005) 810–824.
[116] J.W. Pickering, M.P. Than, L. Cullen, S. Aldous, E. Ter Avest, R. Body, et al., Rapid rule-out of acute myocardial infarction with a single high-sensitivity cardiac troponin T measurement below the limit of detection: a collaborative meta-analysis, Ann. Intern. Med. 166 (2017) 715–724.
[117] A.R. Chapman, K.K. Lee, D.A. McAllister, L. Cullen, J.H. Greenslade, W. Parsonage, et al., Association of high-sensitivity cardiac troponin I concentration with cardiac outcomes in patients with suspected acute coronary syndrome, JAMA 318 (2017) 1913–1924.
[118] P.A. Kavsak, J.T. Neumann, L. Cullen, M. Than, C. Shortt, J.H. Greenslade, et al., Clinical chemistry score versus high-sensitivity cardiac troponin I and T tests alone to identify patients at low or high risk for myocardial infarction or death at presentation to the emergency department, CMAJ 190 (2018) E974–E984.
[119] N. van der Linden, K. Wildi, R. Twerenbold, J.W. Pickering, M. Than, L. Cullen, et al., Combining high-sensitivity cardiac troponin I and cardiac troponin T in the early diagnosis of acute myocardial infarction, Circulation 138 (2018) 989–999.
[120] J.L. Januzzi Jr, W.F. Peacock, A.S. Maisel, C.U. Chae, R.L. Jesse, A.L. Baggish, et al., Measurement of the interleukin family member ST2 in patients with acute dyspnea: results from the PRIDE (Pro-Brain natriuretic peptide investigation of dyspnea in the emergency department) Study, J. Am. Coll. Cardiol. 50 (2007) 607–613.
[121] H. Ohta, T. Tsuji, S. Asai, K. Sasakura, H. Teraoka, K. Kitamura, et al., One-step direct assay for mature-type adrenomedullin with monoclonal antibodies, Clin. Chem. 45 (1999) 244–251.
[122] M.R. Fokkema, Z. Herrmann, F.A.J. Muskiet, J. Moecks, Reference change values for brain natriuretic peptides revisited, Clin. Chem. 52 (2006) 1602–1603.

[123] I. Giuliani, F. Rieunier, C. Larue, J.F. Delagneau, C. Granier, B. Pau, et al., Assay for measurement of intact B-type natriuretic peptide prohormone in blood, Clin. Chem. 52 (2006) 1054−1061.

[124] M.E. Liquori, R.H. Christenson, P.O. Collinson, C.R. Defilippi, Cardiac biomarkers in heart failure, Clin. Biochem. 47 (2014) 327−337.

[125] I.S. Anand, T. Kempf, T.S. Rector, H. Tapken, T. Allhoff, F. Jantzen, et al., Serial measurement of growth-differentiation factor-15 in heart failure: relation to disease severity and prognosis in the Valsartan Heart Failure Trial, Circulation 122 (2010) 1387−1395.

[126] P.A. Kavsak, A. Worster, S.A. Hill, A.R. MacRae, A.S. Jaffe, Analytical comparison of three different versions of a high-sensitivity cardiac troponin I assay over 10years, Clin. Chim. Acta 475 (2017) 51−55.

[127] P.A. Kavsak, C. Ainsworth, L. Clark, P.J. Devereaux, A. Worster, Four different high-sensitivity cardiac troponin assays with important analytical performance differences, Can. J. Cardiol. 35 (2019) 796.e17−796.e18.

[128] J.R. Tate, D.M. Bunk, R.H. Christenson, J.H. Barth, A. Katrukha, J.E. Noble, et al., Evaluation of standardization capability of current cardiac troponin I assays by a correlation study: results of an IFCC pilot project, Clin. Chem. Lab. Med. 53 (2015) 677−690.

[129] A.K. Saenger, R. Beyrau, S. Braun, R. Cooray, A. Dolci, H. Freidank, et al., Multicenter analytical evaluation of a high-sensitivity troponin T assay, Clin. Chim. Acta 412 (2011) 748−754.

[130] A. Rezvanpour, C. Shortt, L. Clark, A. Worster, P.A. Kavsak, Analytical factors to consider when assessing a high-sensitivity cardiac troponin I assay compared to a contemporary assay in clinical studies, Clin. Chim. Acta 429 (2014) 6−7.

[131] A.J. Singer, J. Ardise, J. Gulla, J. Cangro, Point-of-care testing reduces length of stay in emergency department chest pain patients, Ann. Emerg. Med. 45 (2005) 587−591.

[132] V. Palamalai, M.M. Murakami, F.S. Apple, Diagnostic performance of four point of care cardiac troponin I assays to rule in and rule out acute myocardial infarction, Clin. Biochem. 46 (2013) 1631−1635.

[133] P. Venge, C. Ohberg, M. Flodin, B. Lindahl, Early and late outcome prediction of death in the emergency room setting by point-of-care and laboratory assays of cardiac troponin I, Am. Heart J. 160 (2010) 835−841.

[134] J.W. Pickering, J.M. Young, P.M. George, A.S. Watson, S.J. Aldous, R.W. Troughton, et al., Validity of a novel point-of-care troponin assay for single-test rule-out of acute myocardial infarction, JAMA Cardiol. 3 (2018) 1108−1112.

[135] F.S. Apple, M. Panteghini, J. Ravkilde, J. Mair, A.H. Wu, J. Tate, et al., Quality specifications for B-type natriuretic peptide assays, Clin. Chem. 51 (2005) 486−493.

[136] P.A. Kavsak, C.S.P. Lam, A.K. Saenger, A.S. Jaffe, P. Collinson, K. Pulkki, et al., Educational recommendations on selected analytical and clinical aspects of natriuretic peptides with a focus on heart failure: a report from the IFCC Committee on Clinical Applications of Cardiac Bio-Markers, Clin. Chem. 65 (2019) 1221−1227.

[137] A.K. Saenger, R.H. Christenson, Stroke biomarkers: progress and challenges for diagnosis, prognosis, differentiation, and treatment, Clin. Chem. 56 (2010) 21−33.

[138] J.E. Raizman, A. Fuezery, A.K.Y. Tsui, T. Higgins, B.L. Goudreau, L. Clark, et al., Multicenter comparison of imprecision at low concentrations of two regulatory approved high-sensitivity cardiac troponin I assays, Clin. Chim. Acta 486 (2018) 219−220.

Self-assessment questions

1. Which of the following does not occur as part of the pathophysiological events that lead to Type 2 MI?
   a. Atherosclerosis and oxygen supply/demand imbalance
   b. Plaque rupture
   c. Vasospasm or coronary mircovascular dysfunction
   d. Nonatherosclerotic coronary dissection
   e. Oxygen supply/demand imbalance alone
2. For cardiac biomarkers, what is the difference between the 2012 Third Universal MI definition and the 2018 Fourth Universal MI definitions?
   a. There is no difference.
   b. CK-MB is the preferred biomarker in the 2012 definition.
   c. For hs-cTn assays, the sex-specific 99th URLs are recommended in 2018.
   d. For contemporary cardiac troponin assays, a 20% CV at the 99th percentile concentration is acceptable in both the Third and Fourth definitions.
   e. Type 2 MI classification is only in the 2018 definition.
3. According to the ESC/ACC 2000 redefinition of MI, what is the most appropriate cutoff for cardiac troponin?
   a. The functional sensitivity of the assay (concentration at 20% imprecision)
   b. Concentration that differentiates between unstable angina and AMI
   c. Concentration that differentiates between STEMI and NSTEMI
   d. 99th percentile URL from a healthy population
   e. 10% CV concentration
4. Which of the following best describes the clinical sensitivity and specificity of cardiac markers for AMI?
   a. CK-MB has the highest sensitivity and cardiac troponin the highest specificity
   b. Cardiac troponin has the highest sensitivity and CK-MB the highest specificity
   c. CK-MB has the highest sensitivity and specificity
   d. Myoglobin has the highest sensitivity and troponin the highest specificity
   e. Cardiac troponin has the highest sensitivity and specificity
5. Which of the following statements are not true for MI?
   a. There are no longer STEMIs, only Type 1−5 MIs.
   b. Type 4 MI involves MI due to PCI
   c. Type 5 MI involves MI due to CABG
   d. Type 2 MI is an oxygen supply/demand imbalance
   e. Type 1 MI is due to plaque rupture/erosion
6. What is the correct relationship between BNP and NT-proBNP?
   a. Both are derived from proBNP
   b. NT-proBNP is a metabolite of BNP
   c. BNP is a metabolite of NT-proBNP
   d. Both BNP and NT-proBNP are biologically active
   e. Neither BNP nor NT-proBNP are biologically active
7. Which of the following is false regarding the use of BNP and NT-proBNP?
   a. Very low concentrations can differentiate between HF and pulmonary disease
   b. Used to stage the severity of HF
   c. Predicts who will more likely suffer a cardiac event/death
   d. NT-proBNP is a more stable analyte
   e. Guidelines do not endorse measurement of natriuretic peptides for the diagnosis or exclusion of HF in patients presenting with dyspnea
8. Which of the following factors affect the reference interval for BNP and NT-proBNP?
   a. Age
   b. Body mass index
   c. Biological Sex
   d. All of the above
   e. None of the above
9. Which of the following describes the physiologic roles of BNP?
   a. Vasoconstriction, sodium retention
   b. Sodium retention, fluid wastage
   c. Sodium wastage, vasoconstriction
   d. Vasodilation, water retention
   e. Vasodilation, sodium wastage
10. What makes a cardiac troponin assay a “high-sensitivity” assay?
    a. Manufacturer marketing/labeling
    b. An analytical range up to 100,000 ng/L
    c. The ability to measure cardiac troponin concentrations in at least 50% of the healthy population
    d. The 10% CV concentration at the limit of detection
    e. There is no distinction between hs-cTn and current cardiac troponin assays.

Answers

1. b
2. c
3. d
4. e
5. a
6. a
7. e
8. d
9. e
10. c

Chapter 31

Laboratory diagnosis of liver disease

Mark A. Marzinke[1] and D. Robert Dufour[2,3]

[1]Departments of Pathology and Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, United States, [2]Department of Pathology (Emeritus), George Washington University, Washington, DC, United States, [3]Department of Pathology (Adjunct), Northwestern University Feinberg School of Medicine, Chicago, IL, United States

Learning objectives

At the end of this chapter, the reader should be able to:

- Identify biomarkers to assess the synthetic, metabolic, and excretory functions of the liver.
- Describe the biochemical pathway of bilirubin, as well as laboratory methods used in the measurements of conjugated and unconjugated bilirubin fractions.
- Interpret traditional patterns of liver-associated enzymes and injury markers in the background of various types of hepatic disease.
- Differentiate the presentation and laboratory abnormalities encountered in acute and chronic liver diseases.
- Discuss the autoantibodies associated with hepatocellular and cholestatic disease states.

Review of normal liver structure and function

Anatomical features

The liver is a large organ located in the right upper quadrant of the abdomen and weighs approximately 1.2−1.5 kg in a healthy adult. The organ has a dual supply of oxygenated blood, with blood supplied from the portal vein and the hepatic artery. Nutrient-rich blood from the gastrointestinal (GI) tract is carried to the liver via the portal vein and comprises 50%−70% of the oxygenated blood reaching the organ. The hepatic artery transports blood from systemic circulation, and accounts for the remaining 30%−50% of oxygenated blood to the liver. Deoxygenated blood from the liver is transported via the hepatic vein, where it is merged with the inferior vena cava flow and is shunted to the right atrium of the heart. The functional anatomical unit of the liver is the acinus, which is part of the hepatic parenchyma that contains portions of the portal vein, hepatic artery, and bile canaliculus.

There are a number of cell types that are located within the organ, and the cellular architecture of the liver denotes many of its associated functions. In addition to porous endothelial sinusoids, which facilitate blood flow into the organ, the liver contains phagocytic Kupffer cells, which remove antigen−antibody complexes and bacteria from blood. Hepatic stellate cells (Ito cells) exist in both quiescent and activated states. In the former, stellate cells store vitamin A in the form of retinol esters (such as retinyl palmitate); when activated, stellate cells synthesize collagen, and increased collagen production can lead to hepatic scarring and cirrhosis. The predominant cell type within the liver is the hepatocyte. Hepatocytes carry out the synthetic and biotransformative processes attributed to the organ. These processes will be discussed later in this chapter. Between adjacent hepatocytes are bile canaliculi, which are small tubular structures which ultimately fuse to form bile ductules, which merge to form intrahepatic bile ducts. The intrahepatic ducts join to make the common hepatic duct. The hepatic duct joins with the cystic duct from the gallbladder to form the common bile duct. These structures facilitate biliary drainage from the liver into the GI system, and are critical for the excretory functions associated with the organ.

Liver functions

The liver performs a number of functions that may be broadly classified as synthetic, excretory, or metabolic in nature. The liver has an extensive reserve capacity; therefore laboratory tests that assess hepatic function are largely insensitive in the identification of liver damage. Further, biomarkers of liver function may also be affected by conditions other than hepatic injury. An overview of analytes and laboratory tests related to laboratory function will be discussed in this section.

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00031-4

Synthesis

The synthetic capacity of the liver is largely dependent on having sufficient nutrients to carry out essential hepatic functions. The synthetic activities carried out by the liver are dependent not only on the number of hepatocytes, but also the amount of nutrient-rich blood that reaches the organ via the portal vein. Therefore synthetic liver functions are impaired not only in the background of liver disease, but also during states of malnutrition, malabsorption, and portal vein obstruction.

Relevant to its role in synthesis, many enzymes required for glucose and glycogen production are expressed in the liver, as are the building blocks and enzymes required for bile acid and cholesterol biosynthesis and lipoprotein production. The liver also serves as the primary site of production of most plasma proteins (except immunoglobulins), including albumin, transthyretin, ceruloplasmin, α_1-antitrypsin, α-fetoprotein (AFP), and several coagulation proteins. Systemic protein concentrations reflect the balance between synthesis and turnover, and serum or plasma protein patterns observed during liver disease are dependent on both the type and duration of injury, as well as disease state severity.

Because each of the aforementioned proteins varies with regards to half-life, changes in serum or plasma protein concentrations may provide information on the duration of an underlying pathology (Table 31.1). For example, toxic or ischemic hepatitis may result in the global destruction of hepatocytes, thereby decreasing circulating concentrations of clotting factors and leading to increased prothrombin time (PT). The same is not the case with albumin, which is exclusively produced in the liver; due to its long half-life (21 days), hypoalbuminemia is more commonly observed in chronic disease states, such as cirrhosis, autoimmune hepatitis (AIH), and alcoholic hepatitis. Additionally, reduced blood flow from the portal vein deprives the liver of raw materials required for protein synthesis, and is typically associated with decreased protein concentrations. However, due to the reserve capacity of the liver, the majority of individuals with acute hepatic injury do not have abnormal serum or plasma protein concentrations.

TABLE 31.1 Half-life estimates for common hepatically synthesized proteins.

Protein	Half-life
Albumin	19–21 days
Ceruloplasmin	4–5 days
Haptoglobin	2 days
Transferrin	7 days
Transthyretin	2 days
Factor VII	6–8 hours

In addition to liver disease, serum or plasma protein concentrations are influenced by a number of nonhepatic factors, including hormones (estrogens increase ceruloplasmin and HDL production, and cortisol increases transthyretin production), cytokines (interleukin-6 increases acute phase response proteins, such as α_1-antitrypsin, haptoglobin, and ceruloplasmin, while decreasing albumin and a number of other proteins), nutrient intake, and protein-wasting states, such as nephrotic syndrome.

High-throughput automated analyzers typically employ photometric methods for protein measurements, including albumin and total protein. Serum dye-binding assays for albumin include reactions based on bromocresol green or bromocresol purple for the production of a complex that may be measured photometrically. Depending on the specimen source and relative abundance of the protein, antibody-based methods that exploit changes in light scattering may also be employed. A discussion on other methods for the quantification and assessment of protein abnormalities, including protein electrophoretic techniques and immunofixation electrophoresis, is covered in Chapter 22: Proteins: analysis and interpretation of serum, urine, and cerebrospinal fluid, of this book.

Excretion

Compounds in the liver may be biotransformed to facilitate excretion of endogenous or exogenous compounds via the biliary or renal routes. To assess the excretory functions of the liver, laboratory tests include measurement of endogenous compounds, such as bilirubin or bile acids. The former is commonly used in the assessment of hepatic function.

Bilirubin is a molecule that results from the degradation of heme, and is predominantly derived from erythrocyte lysis. Upon cellular lysis, hemoglobin is broken down into heme. Heme is metabolized to the hydrophobic compound biliverdin via heme oxygenase; biliverdin is further processed to form the unconjugated bilirubin molecule via biliverdin reductase. Due to the hydrophobic nature of unconjugated bilirubin, it is loosely bound to albumin when released into the bloodstream to facilitate transport to the liver. Unconjugated bilirubin is taken up by the hepatocytes via a receptor-mediated mechanism. Following cellular uptake, unconjugated bilirubin is rapidly solubilized through the addition of two glucuronic acid moieties via the enzyme uridine diphosphate glucuronosyltransferase (UDP-GT). The more soluble conjugated bilirubin is then excreted into the bile ductules for clearance; this is an energy-dependent, highly efficient

process, and in healthy individuals, very little conjugated bilirubin is detected in the bloodstream. Conjugated bilirubin is transported to the intestine via the common bile duct, where it is converted to urobilinogens via anaerobic bacteria. Urobilinogens are then either reabsorbed for further processing via enterohepatic circulation, or further oxidized in the lower GI tract and cleared in the stool.

However, when conjugated bilirubin in the liver is not excreted, monoconjugated bilirubin may become covalently bound to albumin, forming a stable compound termed biliprotein, or δ-bilirubin. With recovery of liver function, conjugated bilirubin is cleared with a half-life of approximately 24 hours; biliprotein follows the trend of albumin, and is associated with an estimated half-life of 3 weeks. As liver disease improves, therefore, conjugated bilirubin will quickly disappear from systemic circulation, while biliprotein remains elevated over a longer time course. The production, transport, and biotransformation of bilirubin are outlined in Fig. 31.1.

Although we define bilirubin as an unconjugated or conjugated molecule, laboratory tests do not typically report concentrations for each bilirubin fraction. Rather, result reporting reflects traditional methods used for bilirubin detection, and readouts are referred to as total, direct, or indirect bilirubin. The most commonly utilized method for bilirubin measurement is through a reaction with diazotized sulfanilic acid (diazo reagent), which was first described by Ehrlich in 1883. In the early 1900s, it was shown that the addition of an accelerant, such as ethanol (Van den Bergh and Muller), methanol (Evelyn and Malloy), or caffeine (Jendrassik and Grof), solubilized unconjugated bilirubin to facilitate an interaction with the aforementioned diazo reagent. The end product that is formed, referred to as azobilirubin, is measured photometrically. These methods facilitate the measurement of all bilirubin fractions (conjugated, unconjugated, and biliprotein), reflecting total bilirubin concentrations.

Conversely, direct bilirubin assays measure bilirubin mono- and diconjugates, as well as the water-soluble biliprotein, through the direct reaction with the diazo reagent in an alkaline environment. Direct measurements do not require the use of an accelerant. The direct bilirubin reaction typically measures 5% of unconjugated bilirubin and 70%−90% of conjugated bilirubin; the fraction of unconjugated bilirubin that reacts directly with the diazo reagent can be increased by use of surfactants, exposure of serum to ultraviolet light, or alterations in the pH of the reaction or specimen. While the aforementioned methods are most commonly employed in measuring bilirubin fractions, other approaches used to determine bilirubin concentrations include HPLC, direct spectrophotometry, and enzymatic methods. Also, simultaneous determination of unconjugated and conjugated bilirubin fractions may be performed via differential spectrophotometry (spectral shift) with the use of a protein separation layer via a two-slide method; this is performed on the Vitros platform and marketed by Ortho Diagnostics. Notably, this method does not detect biliprotein concentrations.

Increases in unconjugated bilirubin concentrations are associated with not only overproduction of bilirubin, but also increased heme breakdown, as seen during hemolysis, rhabdomyolysis, and in the presence of hematomas. Increased unconjugated bilirubin may also result from impaired bilirubin uptake into the hepatocytes, decreased conjugation via UDP-GT, portal hypertension, and severe hepatic injury. Neonates may present with jaundice and

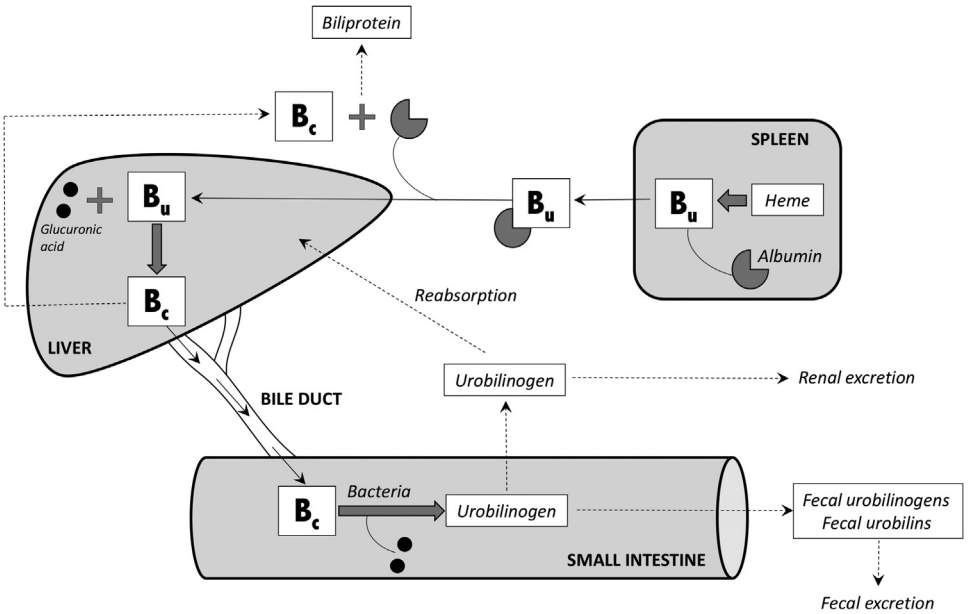


FIGURE 31.1 Bilirubin metabolism. Schematic overview of bilirubin metabolism. B_u: unconjugated bilirubin; B_c: conjugated bilirubin.

increased unconjugated bilirubin concentrations due to decreased conjugation, exposure to drugs that may inhibit bilirubin conjugation, or increased erythrocyte turnover. Neonatal jaundice peaks between 3 days and 2 weeks postpartum. Neonates with sustained jaundice and increased total bilirubin concentrations are treated with phototherapy. Infants are exposed to light at 450 nm to photoisomerize the nonpolar unconjugated bilirubin to produce water-soluble, excretory compounds that may be cleared renally.

There are a number of congenital disorders that can lead to increased concentrations of the unconjugated bilirubin fraction. Crigler–Najjar syndrome is a rare disorder that results from a complete (Type I) or partial (Type II) loss of UDP-GT production. Neonates with Crigler–Najjar Type I present with severe jaundice due to elevated unconjugated bilirubin concentrations in the blood. The hydrophobic, unconjugated bilirubin can cross the blood–brain barrier, resulting in kernicterus, or bilirubin-induced brain damage. Treatment options for infants with Crigler–Najjar syndrome typically involve plasmapheresis or liver transplantation. Less severe disorders that result in increased unconjugated bilirubin concentrations include Lucey–Driscoll syndrome, which is a transient form of familial hyperbilirubinemia, and Gilbert syndrome, a benign form of hyperbilirubinemia associated with mild jaundice.

Conversely, increased conjugated bilirubin concentrations may occur in the background of acute hepatitis, cholestasis, biliary tract obstruction (including biliary atresia), congenital defects, or illnesses that lead to the impairment of bilirubin excretion. Two genetic syndromes that result in increased conjugated bilirubin concentrations are Dubin–Johnson and Rotor syndromes. The former is attributed to an autosomal recessive variant in the canalicular multispecific organic anion transporter gene, resulting in decreased conjugated bilirubin secretion into bile. The clinical presentation varies, but may include mild jaundice or cholestasis; anatomic findings also reveal a hyperpigmented liver due to nonclearance of other pigments, such as lipofuscin. Rotor syndrome is similar in presentation to Dubin–Johnson, but is not associated with liver hyperpigmentation. A summary of hereditary disease states that are associated with increased unconjugated or conjugated bilirubin is presented in Table 31.2.

TABLE 31.2 Increased unconjugated and conjugated bilirubin fractions observed in inherited genetic syndromes.

Syndrome	Unconjugated bilirubin	Conjugated bilirubin
Crigler–Najjar	Elevated	Normal
Dubin–Johnson	Normal	Elevated
Gilbert	Elevated	Normal
Lucey–Driscoll	Elevated	Normal
Rotor	Normal	Elevated

Notes: Increases in unconjugated bilirubin fractions would be observed clinically via increased total bilirubin measurements.

Metabolism

The liver serves as the primary site of metabolism for a number of endogenous and exogenous compounds. The metabolic efficiency of the liver is based not only on the number of hepatocytes, but also on the genetic disposition of metabolizing enzymes. The liver is the primary site of production of a number of metabolizing enzymes, including all of the members of the cytochrome P450 (CYP450) superfamily of enzymes. Many CYP family members are highly polymorphic, resulting in variability in compound metabolism. Since the liver plays a central role in drug metabolism, metabolic tests are useful in assessing hepatic function posttransplantation, as well as in the advanced stages of renal disease.

With regards to endogenous compounds, ammonia metabolism to urea occurs in the liver via the urea cycle. Increased ammonia concentrations (hyperammonemia) can result in hepatic encephalopathy, which is associated with central nervous toxicity and altered mental status; the etiology may be inherited or acquired. Common acquired causes of hyperammonemia include advanced liver and renal disease; increased ammonia concentrations are typical in fulminant hepatitis, Reye's syndrome, and in cirrhotic patients. Additionally, GI bleeding in a cirrhotic patient may lead to increased ammonia production via bacterial metabolism of blood proteins in the colon. However, there is a poor correlation between ammonia concentrations and the severity of hepatic encephalopathy. Ammonia concentrations may also be increased in nonhepatic conditions, including acid–base disorders, infections, or increased protein intake or breakdown. Consequently, ammonia testing is largely nonspecific in assessing the metabolic functions of the liver. Further, ammonia is susceptible to preanalytical factors, such as delayed transportation and separation from red blood cells, as well as prolonged tourniquet use, both of which may falsely elevate ammonia concentrations.

Tests of liver injury

Aforementioned liver function tests are largely insensitive in detecting liver disease because of the reserve capacity of the organ as well as the impact of extrahepatic

pathologies on biomarker concentrations. Therefore markers of liver injury show greater clinical sensitivity and specificity when compared with markers of liver function in identifying hepatic disease. Liver injury markers are helpful both in the recognition and the differential diagnosis of hepatic pathologies, including viral, alcoholic, and toxic hepatitis, as well as biliary tract obstruction and acute liver failure.

The most commonly used biomarkers of hepatic injury or disease are liver-associated enzymes released into systemic circulation. There are a number of mechanisms through which enzymes may be released from a cell into the bloodstream. Enzymes that are associated with various types of liver injury include alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), lactate dehydrogenase (LDH), and gamma-glutamyl transferase (GGT; also known as gamma-glutamyl transpeptidase). Of note, none of these enzymes is expressed exclusively in the liver. A summary of the organ and tissue distribution of each of these enzymes is presented in Table 31.3. Further, within the hepatocyte, the described enzymes differ in terms of their subcellular distribution, as illustrated in Fig. 31.2.

The subcellular localization of these enzymes is an important contributor to the release kinetics of each macromolecule and the clinical correlation with various types of liver injury. Because all cells try to maintain their cellular machinery, macromolecules localized within the cytoplasm are primarily released after irreversible cellular injury. However, membrane-bound enzymes may be released during reversible cellular damage, such as in the background of cholestatic conditions. Induction of enzyme synthesis within cells also increases serum or plasma levels of inducible enzymes, such as GGT. Laboratory assessment of hepatic enzymes typically involves the measurement of enzyme activity as opposed to concentrations; thus enzyme results are reported as units per liter (U/L). The specific patterns of these enzymes in the background of various disease states will be discussed later in this chapter.

TABLE 31.3 Characteristics of liver-associated enzymes.

Enzyme	Tissue/organ sites of expression	Subcellular distribution	Half-life
AST	Liver, heart, muscle, and kidney	Cytoplasm and mitochondria	cAST: 16–18 hours mAST: 10 days
ALT	Liver and kidney	Cytoplasm	40–48 hours
LDH	Liver, heart, RBC, muscle, kidney, and WBC	Cytoplasm	4–6 hours
ALP	Liver, bone, intestine, and placenta	Membrane	24 hours
GGT	Liver, pancreas, kidney, and prostate	Membrane and microsomes	4 days (longer with chronic alcohol abuse)

ALP, Alkaline phosphatase; *ALT*, alanine aminotransferase; *LDH*, lactate dehydrogenase; *AST*, aspartate aminotransferase; *cAST*, cytoplasmic aspartate aminotransferase; *GGT*, gamma-glutamyl transferase (transpeptidase); *mAST*, mitochondrial aspartate aminotransferase; *RBC*, red blood cells; *WBC*, white blood cells.

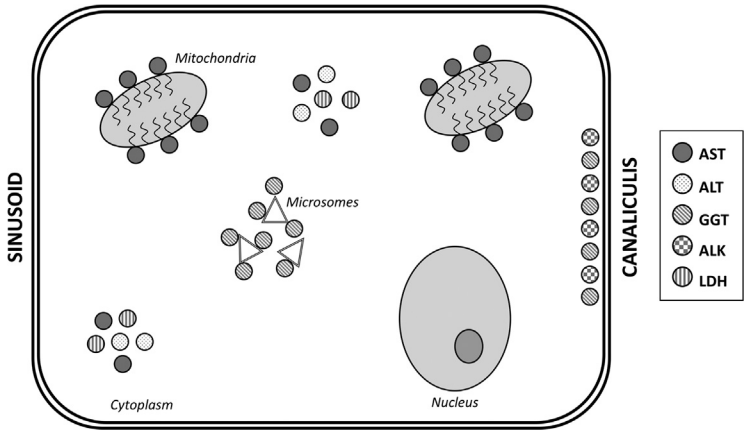


FIGURE 31.2 Subcellular distribution of liver-associated enzymes. The major hepatocyte enzymes have different locations within the cell, which explains the patterns of abnormalities encountered in various forms of liver disease. *ALP*, Alkaline phosphatase; *ALT*, alanine aminotransferase; *AST*, aspartate aminotransferase; *GGT*, gamma-glutamyl transferase; *LDH*, lactate dehydrogenase.

Liver-injury markers

Aspartate aminotransferase

Physiologically, the enzyme AST catalyzes the reversible transfer of an amino group from glutamate to oxaloacetate for the generation of aspartate. Consequently, the enzyme was formerly referred to as serum glutamic-oxaloacetic transaminase. Both AST- and ALT-mediated enzymatic reactions require a pyridoxal phosphate (vitamin B_6) cofactor. Pyridoxine deficiency may occur during states of chronic alcohol abuse, malnutrition, or in the presence of pyridoxine binders. Thus several commercial vendors offer assays that are supplemented with pyridoxal phosphate to ensure sufficient cofactor concentrations. As indicated in Fig. 31.2, the subcellular distribution of AST is in both the hepatocyte cytoplasm as well as within the mitochondria; the cytoplasmic portion is the predominant form of AST released posthepatic injury. Alcohol abuse has been shown to induce the expression and release of mitochondrial AST. However, even in alcoholic liver disease, less than 10%–20% of total AST is the mitochondrial isoenzyme. Notably, the half-lives of the cytoplasmic and mitochondrial AST isoenzymes differ; the half-life of cytoplasmic AST is 16–18 hours, while mitochondrial AST is approximately 10 days. Consequently, in the background of chronic alcohol abuse, AST activity measurements may remain elevated above the upper reference limit of normal for a longer period of time than ALT.

Alanine aminotransferase

The physiological function of ALT is similar to that of AST. ALT catalyzes the reversible transfer of glutamate to pyruvate, generating alanine; historically, ALT was referred to as serum glutamic-pyruvic transaminase. Like AST, ALT is also measured in the clinical laboratory via the determination of enzyme activity, and ALT activity is more sensitive to vitamin B_6 status than AST. ALT is the liver enzyme associated with the fewest other tissue or organ systems, and is primarily located within the cytoplasm of hepatocytes. While it is less abundant in the hepatocyte than AST, it has a longer half-life, estimated to be 40–48 hours. Consequently, following acute liver injury, while AST levels may rise first, ALT activity measurements are elevated above the upper reference limit of normal for a longer period of time. ALT is a key liver-injury marker that is used in the differential diagnosis and monitoring of liver injury or disease. Further, AST/ALT activity ratios can be helpful in the differential diagnosis of hepatic inflammation etiology. This will be discussed in further detail later in this chapter.

Lactate dehydrogenase

As illustrated in Table 31.3, LDH is widely distributed in tissue, with high protein expression in erythrocytes, as well as the heart, liver, muscles, and kidney. Further, there are multiple isoenzymes that exist, which can be resolved electrophoretically. LDH of liver origin has a short half-life of 4–6 hours, and is found at lower levels than other liver-associated enzymes following hepatic damage. While LDH is increased in a number of disease states, including megaloblastic anemia, shock, myocardial or pulmonary infarction, or hemolytic anemia, liver-related causes for LDH activity elevation are primarily toxic or ischemic hepatitis (when patients present soon after onset of injury, allowing detection of the short-lived liver isoenzyme).

Alkaline phosphatase

Although ALP is expressed in a number of tissues, circulating concentrations typically reflect ALP derived from the liver or bone. In the liver, ALP is a membrane-associated enzyme that is released in response to canalicular damage. Damage to the canalicular surface of the hepatocyte may be caused by either drainage obstruction or direct damage to the canaliculi via drugs or viral infections. In response, ALP levels begin to rise within 24–48 hours, and the enzyme has a half-life of 24 hours. ALP is measured by determining enzyme activity; notably, enzymatic activity requires divalent cation cofactors, such as zinc or magnesium. States of clinical zinc deficiency or collection in tubes containing chelating agents (such as EDTA) can result in artificially decreased ALP activity measurement results. While ALP derived from the bone and liver are encoded by the same gene, they are differentially posttranslationally modified and differ with respect to their carbohydrate content. Isoenzymes may be differentiated via heat fractionation, electrophoresis, isoelectric focusing, or antibody-based approaches. With regards to heat fractionation, the lability of various ALP isoenzymes is exploited, as bone ALP is the most heat labile, liver ALP is less so, and another form of ALP, placental ALP, is heat stable. For laboratories that discriminate ALP isoenzymes, an agarose electrophoretic approach is most commonly employed.

Gamma-glutamyl transferase

GGT is a glycoprotein enzyme that is expressed on the canalicular membrane as well as within hepatic microsomes. Similar to ALP, GGT levels will increase in response to canalicular damage. GGT measurements are typically used in conjunction with ALP to assess bile duct obstruction or other forms of ductule damage. GGT activity measurements may also be increased following exposure to microsomal enzyme-inducing drugs, such as ethanol, phenobarbital, phenytoin, and carbamazepine. Regarding ethanol abuse, GGT activity measurements are less likely to be increased in binge drinkers or those consuming less than 2–3 drinks per day, as compared with

chronic, heavy drinkers. Once released into the bloodstream, the half-life of GGT is several days; further, when released from the microsomes in the background of chronic alcohol abuse, GGT levels may be elevated for several weeks.

Other liver biomarkers

In addition to the aforementioned liver-associated enzymes, there are several other biomarkers that can be utilized to assess and monitor liver injury, including hepatic cancers and autoimmune disorders.

Alpha-fetoprotein

AFP is a major plasma protein expressed at high concentrations during fetal development; however, concentrations decrease drastically after birth. When hepatocytes regenerate or proliferate, AFP concentrations may increase. Elevated AFP concentrations may be observed in cirrhotic patients, during a flare-up of chronic hepatitis, or during the recovery phase postacute hepatitis. Additionally, AFP is commonly used as a tumor marker for hepatocellular carcinoma (HCC) or germ cell (nonseminoma) carcinoma. With regards to HCC, AFP may be used as a prognostic indicator of survival as well as in monitoring disease progression. AFP can be further stratified based on protein binding to the lectin *Lens culinaris* agglutinin. One form of AFP, AFP-L3, binds strongly to the aforementioned lectin and is associated with malignant tumors, such as HCC. Consequently, an AFP-L3% test (which is the percentage of AFP that is the L3 isoform) at a cutoff of 10% has shown utility as a tumor marker for HCC.

Des-gamma-carboxy prothrombin

Des-gamma-carboxy prothrombin (DCP), also known as PIVKA-II, is a form of factor II produced in the absence of vitamin K, and is therefore an unmodified form of the protein. In most individuals with liver disease, PIVKA-II is undetectable, but in acute hepatitis and HCC, PIVKA-II concentrations may be increased, often in individuals with minimal or no increase in AFP concentrations. Conversely, PIVKA-II is less commonly elevated in the background of chronic hepatitis or cirrhosis. PIVKA-II is sensitive to vitamin K antagonists, and may be increased in individuals with vitamin K deficiency or on warfarin therapy. In recent years, this biomarker has gained interest as a potential tumor marker for HCC.

Autoantibodies

There are a number of liver diseases that are characterized by the presence of autoantibodies. AIH is a form of chronic hepatitis that is associated with long-term hepatic inflammation and liver damage. While clinical symptoms may be similar to those observed in acute hepatitis, physical findings show a high prevalence of hepatomegaly and jaundice in patients with AIH; ascites is also seen in more severe cases. Most cases of AIH are associated with increased concentrations of immunoglobulins, particularly IgG, as well as the presence of antinuclear antibodies, antismooth muscle antibodies, and antiliver−kidney microsomal type 1 antibodies. Patients with AIH are typically treated with immunosuppressive therapies, such as prednisone and azathioprine. Autoantibodies associated with cholestatic diseases will be discussed later in this chapter.

Assessment of liver disease

Liver diseases are typically stratified as being acute or chronic in their presentation. Etiologies of acute and chronic liver damage will be discussed in this chapter, as well as the role that laboratory testing plays in the differential diagnosis of hepatic injury.

Clinical manifestations of liver disease

There are a number of clinical manifestations that are suggestive of hepatic injury. Jaundice, which is a yellowing of the skin due to excess bilirubin in the bloodstream, is a common manifestation that is typically associated with acute presentations of obstruction of the bile ducts or hepatocyte necrosis. Consequently, urine is dark and stool is light in color. However, children and a percentage of adults with acute hepatitis do not become jaundiced, which may lead to delayed workup of acute liver injury. Jaundice is also a late finding in persons with cirrhosis, In addition to hepatic injury, jaundice may also present in the background of increased intravascular hemolysis, hematomas, rhabdomyolysis, or in previously discussed inborn errors of metabolism. Dysregulation of hemostasis and portal hypertension, which may be caused by cirrhosis and can result in decreased blood flow to the liver, are additional features that may be observed in patients with liver disease. The presence and extent of these manifestations are dependent on the underlying pathology.

Acute liver disease

Acute hepatitis involves injury to the hepatocytes that may be viral/immunological, ischemic, or drug-induced in nature. One common cause of hepatitis, particularly in developing countries, is viral infection, which results in immunological responses, hepatic inflammation, and release of liver-associated enzymes. Ethanol also elicits an immunological response on the liver, resulting in hepatocyte injury. Ischemic or toxic hepatitis can lead to direct

liver damage and hepatocyte necrosis. For example, an acetaminophen overdose can lead to hepatic accumulation of *N*-acetyl-*p*-benzoquinone imine (NAPQI), resulting in cellular death. Ischemic hepatitis may be caused by hepatic hypoperfusion. Of note, ischemic and toxic forms of hepatitis typically impact markers of liver function as well as liver injury. Regardless, all forms of acute hepatitis are associated with irreversible damage to hepatocytes and the subsequent release of biomarkers into systemic circulation. Hepatocyte death may be apoptotic or necrotic in nature; the former being the more common mechanism of cell death.

Clinical presentation and patterns of injury associated with immunologically mediated acute hepatitis

Acute hepatitis that is ethanol-induced or caused by viral infections follows a similar trajectory in terms of clinical presentation. Individuals typically experience an asymptomatic incubation period, followed by a prodromal (early onset) period of nonspecific symptoms, which may include fever, malaise, vomiting, and/or nausea. Viral infections in older children and adults may also result in increased bilirubin concentrations and jaundice, and present with liver tenderness, dark-colored urine, and lightly colored stool. Because liver damage is gradual and prolonged in these disease states, the release of cytoplasmic enzymes shows a gradual increase, reaching a plateau, before demonstrating a gradual decline.

AST and ALT activity measurements of 200 and 300 U/L, respectively, have clinical sensitivity and specificity of 90% in supporting the diagnosis of acute viral hepatitis. These values correlate with approximately five and eight times the upper reference limits of normal, respectively. With viral infections, AST and ALT may reach 10−50 times the upper reference limits of normal throughout the disease course. While AST measurements rise first in viral hepatitis, ALT measurements are typically higher at the time of clinical presentation. This is due to the longer half-life of ALT as compared with AST. Aminotransferase ratios are also useful in the differential diagnosis of acute liver injury, and AST/ALT ratios in viral infections are typically <1 for the aforementioned reasons. In the background of viral hepatitis, ALP activity measurements may also be mildly elevated. Further, increases in aminotransferases typically precede increases in bilirubin concentrations, and may remain elevated for several weeks before gradually decreasing.

Alcoholic hepatitis occurs in only about 10% of those who abuse alcohol; however, some studies indicate that immune responses to acetaldehyde-altered proteins may be involved in the pathogenesis of hepatic injury. There are several laboratory findings that distinguish alcoholic hepatitis from other forms of liver injury. AST and ALT measurements are seldom over 10 times the upper reference limits of normal, and unlike viral hepatitis, AST activity measurements are usually greater than ALT (often more than two-fold higher). Studies have demonstrated that a percentage of AST released during alcoholic hepatitis is mitochondrial AST; *in vitro* studies have demonstrated that alcohol can stimulate the release of the longer half-life mitochondrial AST from the hepatocyte. Additionally, unlike viral hepatitis, alcoholic hepatitis is more commonly associated with increased levels of ALP, and usually GGT. In chronic alcoholics, GGT measurements may be elevated for several weeks; however, there is no correlation between the extent of GGT activity elevation and disease severity.

In cases of jaundice, which may occur due to hepatitis or biliary tract obstruction, conjugated bilirubin predominates, with direct bilirubin accounting for 50%−80% of total bilirubin concentrations. The glucuronidated, water-soluble bilirubin fraction can be renally cleared, resulting in the presence of bilirubin in the urine. Early in hepatitis, biliprotein typically comprises 15%−20% of total bilirubin; with resolution, the percentage increases to approach 100%. The declining presence of bilirubin from urine is a positive sign of recovery posthepatic injury. However, severe liver damage and cellular death can lead to decreased bilirubin conjugation, resulting in increased concentrations of the unconjugated bilirubin fraction. Other liver function tests, including those that assess the synthetic capacity of the liver, are typically within normal reference intervals in the background of viral or alcoholic hepatitis; prolonged PT is a poor prognostic marker with both of these causes of liver injury. Patterns of hepatic function and injury markers in various types of acute hepatitis are summarized in Table 31.4.

Causes of viral acute hepatitis

Hepatitis A

Hepatitis A virus (HAV) is a single-stranded RNA picornavirus that is transmitted via the fecal−oral route, and transmission may occur through the ingestion of contaminated food or water. Previously, hepatitis A was a common cause of hepatic injury in the United States; however, since the licensure of the hepatitis A vaccine in 1995, there has been a marked reduction in the incidence of new infections. However, outbreaks in developing countries, particularly in areas with inadequate water sanitation resources, still remain a public health concern. Postinfection, the virus undergoes an incubation period ranging from 2 to 6 weeks, before the presentation of the aforementioned nonspecific symptoms

TABLE 31.4 Laboratory patterns associated with viral, alcoholic, and toxic/ischemic forms of acute hepatitis.

Feature	Viral	Alcoholic	Toxic/ischemic
% Chronic	Variable	5%–10%	None
% Fatal	<3% (except D)	5%–10%	10%
AST/ALT ratio (at diagnosis)	<1	Usually >2	>1 (transient)
Peak AST (× normal)	10–100	1–10	>100
LDH (× normal)	1–2	1–2	10–40
Peak bilirubin, milligram per deciliter (micromole per liter)	5–20 (85–340)	3–20 (51–340)	Usually <5 (85)
Prothrombin time	Normal	Normal	Usually >15 s

ALT, Alanine aminotransferase; *AST*, aspartate aminotransferase; *LDH*, lactate dehydrogenase.

(including vomiting, nausea, and fatigue). Infected adults typically present with jaundice, but infected children may not exhibit any clinical symptoms. Further, approximately one quarter of individuals are noninfectious at clinical presentation. As demonstrated in Table 31.4, markers of liver function or injury, in particular ALT, may be elevated.

Serologically, the presence of IgM anti-HAV is diagnostic of an acute infection in the proper clinical setting. The IgM form of anti-HAV is typically present at clinical onset of acute hepatitis, and may be detectable via laboratory assays for up to 6 months postinfection. Conversely, total (mostly IgG) antibody concentrations persist following infection or immunization. In the vast majority of cases, hepatocyte inflammation and injury resolve in a matter of weeks; however, if a patient infected with chronic hepatitis C becomes infected with HAV, there is an increased risk for hepatic necrosis and liver failure.

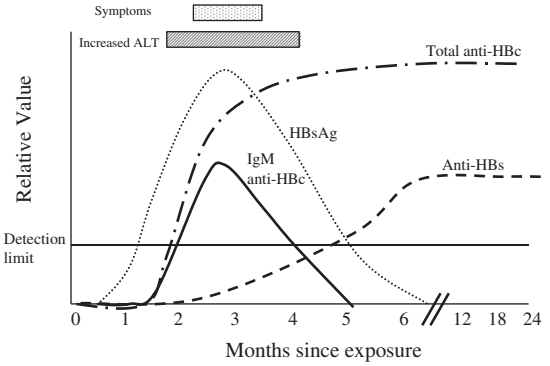


FIGURE 31.3 Clinical course of viral hepatitis B. After a period with no detectable viral antigens, hepatitis B surface antigen appears in the circulation. With the onset of immune response to the virus, liver damage occurs, as manifested by the hatched bar indicating time of elevation of enzymes. After an incubation period, jaundice typically develops (J). Following resolution of the acute damage, jaundice resolves and, eventually, enzymes return to normal, and hepatitis B surface antigen is cleared with appearance of anti-HBs. Antibodies to hepatitis B antigens persist for the life of the patient. *Reproduced with permission from W. Clarke (Ed.), Professional Practice in Clinical Chemistry, second ed., AACC Press, Washington, DC, 2011.*

Hepatitis B

Hepatitis B virus (HBV) is a double-stranded DNA hepadnavirus. Recent estimates indicate that up to one-third of the global population has been infected with hepatitis B, with roughly one-tenth of these cases leading to chronic viral infection. However, as with HAV, the frequency of infection has declined markedly following the implementation of vaccination, although there has been a recent increase in incidence with the current opioid crisis. HBV is transmitted via infected bodily fluids, and routes of contact require sexual contact, mother-to-child (vertical transmission), or blood-based exposures.

The incubation period of the DNA virus ranges from 3 to 6 months, and many of the previously described symptoms may present in an infected individual; 30%–50% of affected individuals will develop jaundice. Further, a small percentage of those infected (<2%) may have significant hepatic injury, resulting in fulminant hepatic failure. Unlike HAV, HBV infection can persist, and an individual may become chronically infected. With a full immune response, following an acute infection, the virus is typically "cleared." However, approximately 2% of immune-competent older children and adults fail to clear the virus and develop chronic HBV; this figure rises to 10%–20% in immunocompromised individuals (including HIV-infected individuals or those using immunosuppressive drugs) and >90% in neonates.

The typical time course of acute HBV infection is illustrated in Fig. 31.3. Both the hepatitis B surface antigen (HBsAg) as well as the hepatitis B core antibody, IgM (anti-HBc, IgM) are expressed during the early stages of an active HBV infection; the HBsAg is

produced at high concentrations by the double-stranded virus, and the anti-HBc, IgM is generated in response to the infection. These biomarkers are detected within the first 6 months of an acute infection. The hepatitis B e antigen (HBeAg) is also a marker of active infection, and detection of this antigen correlates with HBV infectivity when strains produce this antigen (although not all do). The first viral marker of recovery is loss of HBeAg and appearance of anti-HBe. Loss of HBsAg and development of anti-HBs typically take 3–6 months, and is considered evidence of viral clearance; however, low levels of viral DNA usually persist in the liver. Detection of anti-HBs and the total anti-HBc antibodies can indicate a past-infection or immunization (in which case, only anti-HBs will be present). A full discussion of the hepatitis B life-cycle, HBV genotypes, and issues regarding HBV mutations may be found in Chapter 54: Infectious diseases, of this book.

Hepatitis C

Hepatitis C virus (HCV), an RNA flavivirus, causes about 15%–20% of acute viral hepatitis infections in the United States. Intravenous drug use and transfusion of infected blood are routes of infection; however, screening of blood products prior to transfusion has dramatically reduced this route of transmission. As with HBV, there has been a recent rise in incidence correlated with injected opioid use. HCV can also be transmitted from mother to child, but the incidence of this is much lower when compared with HBV. Acute HCV infection has a variable incubation period, with an average period of 6–12 weeks. Elevations in aminotransferases are typically observed in the prodromal incubation period. Similar to HBV, HCV infections may persist. Chronic HCV infections may occur in up to 85% of those infected with the virus; roughly, one-fifth of these individuals may experience cirrhosis, hepatic failure, HCC, or cholangiocarcinoma. Those who are infected through injection drug use may also be infected with HIV and/or HBV. Coinfection can lead to increased hepatic-related morbidity and mortality and can exacerbate viral hepatitis disease progression and liver damage.

With regards to serological testing, IgG antibodies against HCV (anti-HCV) represent the most widely used test to screen for HCV exposure. Currently, second and third generation assays exist (third generation tests can detect HCV antibodies 9 weeks after exposure). Further, the CDC provides recommendations and guidelines, describing signal-to-cutoff ratios for commercial immunoassays. In addition to anti-HCV testing, molecular testing, including HCV RNA analysis, may be used to detect an active infection. Many clinical laboratories now implement reflex quantitative HCV RNA testing, as HCV viral load results can assist in directing treatment.

Hepatitis D

Hepatitis D virus (HDV) is a defective RNA virus that can replicate only in the presence of the HBsAg. It is primarily transmitted via blood. While it is endemic in some parts of the world, it has become less common due to the decline in new cases of HBV infections. In individuals simultaneously infected with both HBV and HDV, acute hepatitis is more severe, and liver failure is more likely. In a person with chronic HBV who subsequently becomes infected with HDV, there may be a flare in disease activity that mimics acute hepatitis. Diagnosis is usually determined through the detection of anti-HDV antibodies, but tests for the HDV antigen and nucleic acid testing may also be used.

Hepatitis E

Hepatitis E virus (HEV) is a single-stranded RNA virus. While not common in developed countries, HEV infections are still endemic in developing and rural areas. HEV is transmitted enterically, and as such, clinical manifestations are similar to HAV. However, HEV can lead to significant hepatic failure and increased mortality during pregnancy. HEV has also been found to cause chronic hepatitis in immunocompromised individuals. Currently, the only commercially available tests for HEV are total and IgM anti-HEV.

Toxic and ischemic hepatitis

Toxic and ischemic hepatitis are associated with abrupt changes in liver function and extensive cell death, resulting in significant elevations in cytoplasmic enzymes as well as changes in liver function tests. Laboratory patterns in comparison with viral and alcoholic hepatitis are summarized in Table 31.4. Toxic hepatitis, as discussed earlier, involves direct damage to the liver, most commonly caused by drugs. The most common cause of toxic hepatitis is an overdose of acetaminophen. Following accumulation of the toxic NAPQI metabolite, cellular necrosis occurs; clotting factors are depleted, leading to increased PT. Further, 24–48-hours postingestion, elevations in AST, ALT, and LDH activities are observed. Enzyme activities may exceed 100 times the upper reference limit of normal. Bilirubin is usually minimally elevated, and may not reach peak concentrations until a few days after peak abnormalities in enzymes or PT occur (Fig. 31.4). Prolonged elevation of PT, which indicates impaired synthetic capacity of the liver, is indicative of a poor prognosis.

From a laboratory perspective, ischemic hepatitis is similar to toxic hepatitis, as both are associated with massive cellular necrosis. Ischemic hepatitis, also known as "shock liver" is the most common cause of acute hepatitis

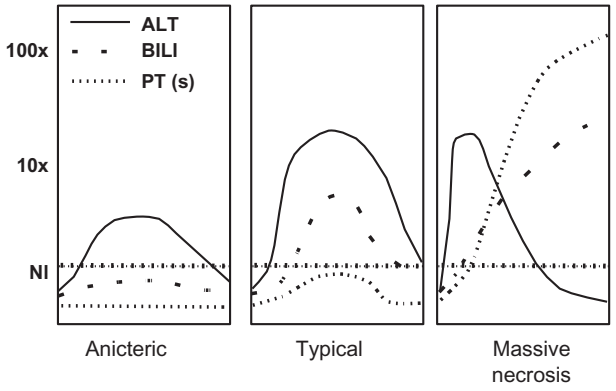


FIGURE 31.4 Results of blood tests in various severity of hepatitis. The patterns of changes in bilirubin (dashed line), alanine aminotransferase (solid line), and prothrombin time (dotted line) are shown for the three major types of hepatitis. In patients with anicteric disease, serum aminotransferases rise slightly, but liver function remains normal. With typical icteric hepatitis, patients have more marked enzyme increases and become jaundiced with increased serum bilirubin; prothrombin time usually remains normal. With massive hepatic necrosis, the loss of all functioning cells causes transaminases to ultimately return to normal limits; however, liver function continues to worsen; transplantation is the only effective treatment. *Reproduced with permission from W. Clarke (Ed.), Professional Practice in Clinical Chemistry, second ed., AACC Press, Washington, DC, 2011.*

in hospitalized patients. It can also occur as a part of the spectrum of hepatic sickle cell crisis or as a consequence of hepatic artery thrombosis in patients with compromised portal circulation. Schematics of ALT, bilirubin, and PT profiles in the background of minimal, nonjaundiced injury, moderate hepatic injury (typically associated with jaundice), and massive cellular necrosis are illustrated in Fig. 31.4.

Biliary tract obstruction

Obstruction is a less common cause of acute jaundice in children and young adults but becomes more prevalent in older individuals. The importance of distinguishing obstructive jaundice from hepatitis is that, while hepatitis is treated conservatively, obstruction is often due to potentially treatable causes that can benefit from immediate intervention. Complete obstruction of both hepatic ducts or extensive intrahepatic obstruction are required to produce obstructive jaundice; the most common causes are gallstones and tumors of the pancreas in adults and congenital biliary atresia in children. With incomplete obstruction, patients often remain asymptomatic and do not present with jaundice.

In the early stages of obstruction, particularly in the background of gallstones, laboratory features are similar to those seen during hepatitis. A comparison between hepatitis and biliary obstruction in terms of laboratory values is summarized in Table 31.5. Initially, increased ductal pressure causes hepatic damage, leading to increased AST and ALT activity measurements. However, as obstruction persists, both increased liver synthesis and bile acid-mediated solubilization of fragments of the canalicular membranes lead to increased release of ALP and GGT. Even with continued obstruction, AST and ALT typically fall, with AST returning to normal within 8–10 days. The degree of elevation in bilirubin and ALP are similar in intrahepatic and extrahepatic causes of obstruction. Bilirubin is mostly direct reacting (50%–80%), similar to the percentage observed in the background of acute hepatitis.

Chronic liver disease

Chronic liver diseases can involve prolonged inflammation or damage to the hepatocytes, as well as longer-term damage, including fibrosis and cirrhosis. In chronic liver diseases, jaundice is either rare or is a late finding, occurring usually after a clinical diagnosis has been made. Common causes of chronic liver disease include persistent viral infections with HBV or HCV, as well as nonalcoholic fatty liver disease (NAFLD), AIH, hemochromatosis, and Wilson's disease. As scarring progresses to cirrhosis, not only are liver injury markers impacted, but synthetic, metabolic, and excretory functions of the liver are affected.

Chronic hepatitis

Chronic hepatitis is the most common cause of chronic liver disease, and is defined as prolonged inflammation and hepatocyte damage typically lasting >6 months; persistent damage can lead to necroinflammation and may or may not be accompanied by hepatic scarring (fibrosis). Individuals with chronic hepatitis may be asymptomatic or present with nonspecific symptoms, such as fatigue or weakness. Chronic hepatitis may be identified by sustained mild elevations of ALT (commonly) or AST activity for 6 months in the absence of other etiologies; elevations in aminotransferase activity average two times the upper reference limit of normal. ALP and GGT activity are typically within reference intervals, as is bilirubin.

Elevated ALT activity measurements correlate more closely with inflammation than fibrosis. Currently, liver biopsies are the standard for the identification of hepatic fibrosis and disease progression. Biomarkers for collagen formation, including laminin, hyaluronate, and collagen fragments, may be observed following fibrosis, but there is weak correlation between biomarker concentrations and disease severity. Increases in aminotransferase activity and compromised synthetic capacity of the liver can be observed during progression to cirrhosis, but these markers alone do not yield sufficient clinical sensitivity for

TABLE 31.5 Differences in presentation, clinical, and laboratory findings between hepatitis and biliary obstruction.

Feature	Hepatitis	Obstruction
Symptoms	Painless jaundice	Jaundice, may be accompanied by pain or silent
% Conjugated bilirubin	50–80	50–80
AST and ALT	>10 times normal (except alcoholic)	<10 times normal (but may transiently increase >10 times)
Alkaline phosphatase	<3 times normal	<3 times normal early and >5–10 times normal later
Bile duct imaging	Normal	Dilated

ALT, Alanine aminotransferase; *AST*, aspartate aminotransferase.

diagnosis. Combinatorial testing assessing liver-injury and proinflammatory markers to predict severity of fibrosis have been patented to predict liver scarring. The FibroTest (referred to in the United States as FibroSure) is a multiplexed panel, including total bilirubin, GGT, haptoglobin, apolipoprotein A_1, α_2-macroglobulin, and demographic information (age and sex) marketed for the diagnosis of fibrosis. Studies comparing the serum biomarker test to liver biopsies demonstrate approximately 25% discordance between the two approaches. Further, when correlations have been performed between aberrant laboratory values and the extent of fibrosis, there is typically a large group of patients who have "indeterminate" results, and the ability of these markers to work in all causes of liver injury has not been demonstrated.

Chronic viral hepatitis

In chronic viral infections, viral markers (HBsAg or HBV DNA, and HCV RNA) are more reliable than liver injury enzymes in the evaluation of chronic hepatitis, as well as downstream initiation and monitoring of therapy. Sustained detection of HBsAg indicates chronic HBV infection. If HBsAg is positive, HBV DNA quantitation is also used to confirm continued replication of virus, and to serve as a baseline value for therapy. Antiviral therapy is commonly used for the treatment of chronic HBV infection, with the goals of suppressing viral propagation and mitigating hepatic injury, which are assessed by decreased HBV DNA levels (with a goal of undetectable HBV DNA) and reduction in ALT activity, respectively. In HBeAg positive individuals, the goal of treatment is to first clear circulating virus from systemic circulation, then clear HBeAg to facilitate the production of anti-HBe. Rarely, individuals will also lose HBsAg and develop anti-HBs. If anti-HBe does develop, treatment should continue for at least 6 months, but treatment may be continued indefinitely if HBV DNA becomes undetectable without loss of HBeAg (or in those individuals who were HBeAg negative before treatment).

Initial assessment of chronic HCV infection includes laboratory testing for anti-HCV. If anti-HCV is positive, HCV RNA quantitation is usually performed to confirm infection. Prior to the initiation of therapy, HCV genotyping is performed to direct treatment modalities and duration. The typical treatment regimen for chronic HCV infection most commonly involves a combination of direct antiviral agents; other drugs (such as ribavirin and ritonavir) may be used depending on HCV genotype. The goal of treatment is the loss of HCV RNA, which persists after treatment is stopped. While most patients are monitored by checking HCV viral loads after 4 weeks on treatment, current treatments are so highly effective that the value of this testing has been questioned. HCV RNA testing is repeated 3–6 months posttreatment. If HCV RNA remains undetectable, the patient is said to have had a sustained virologic response, and there is <1% chance that viremia will recur.

Other causes of chronic hepatitis

NAFLD and nonalcoholic steatohepatitis (NASH) are associated with fat accumulation and hepatic inflammation; both are associated with obesity, diabetes, hypertriglyceridemia, and low HDL-cholesterol concentrations in the absence of significant alcohol use. Laboratory abnormalities are mainly limited to increased aminotransferase activity. While NAFLD is more prevalent than NASH, the latter is more likely to lead to extensive liver scarring and cirrhosis.

AIH, Wilson's disease, Budd–Chiari syndrome, α_1-antitrypsin deficiency, veno-occlusive disease, and hemochromatosis are additional causes of chronic hepatitis and prolonged hepatic injury. AIH is associated not only with increased aminotransferase activity and the presence of previously described autoantibodies, but hypoalbuminemia may also be observed. Wilson's disease, an inherited disorder that leads to copper accumulation in the liver (as well as the brain and tissue), is associated not only with hepatic inflammation and scarring, but in acute

presentations patients may also present with hemolytic anemia and tubular necrosis due to release of excess copper from the liver.

Cirrhosis

Cirrhosis is the end stage of scar formation, usually as a result of chronic hepatitis or chronic bile duct obstruction. Early in the transition to cirrhosis, there are no clinical or laboratory features that allow its recognition. As scarring progresses, protein vein flow decreases, causing portal hypertension. This may become apparent clinically through the development of ascites, bleeding from esophageal varices, or hepatic encephalopathy. Laboratory features that suggest progression to cirrhosis include decreased hepatic proteins, (low albumin and prolonged PT), polyclonal gammopathy (especially with increased IgA, causing $\beta-\gamma$ bridging on serum protein electrophoresis), reversal of AST/ALT ratio to >1, and a fall in platelet count. Increased total bilirubin (mainly indirect reacting) is a later finding in persons with cirrhosis. As liver failure develops, albumin, PT, and creatinine have been found to be of prognostic significance, and are used to calculate the model of end stage liver disease (MELD) score, calculated as:

$$\begin{aligned} &3.8 + \ln \text{bilirubin } (\text{mg/dL}) \\ &\quad + 11.2 \ln \text{International Normalized Ratio} \\ &\quad + 9.6 \ln \text{creatinine } (\text{mg/dL}) \\ &\quad + 6.4 \text{ etiology score} \\ &\quad (0 \text{ if alcohol or obstruction; } 1 \text{ for all other causes}) \end{aligned}$$

which is the major criterion for consideration for liver transplantation. MELD scores of >30 are associated with very high mortality and give preference for liver transplantation.

Chronic cholestatic disorders

Chronic cholestasis is defined by one or more of a constellation of findings, which may include increases in canalicular enzymes (such as ALP), dilated bile ducts on imaging studies, skin itching (pruritus), elevated cholesterol (due to accumulation of lipoprotein X), and deficiency of fat-soluble vitamins. One form of cholestatic disease is primary biliary cholangitis (PBC), which is a disease that causes the inflammation and subsequent destruction of the intrahepatic bile ducts, leading to a buildup of bile in the liver and subsequent hepatic scarring. PBC is typically a progressive disease that can lead to liver failure and require transplantation; it is also associated with an increased risk of HCC. There is higher incidence of PBC in middle-aged women. Clinically, PBC is often associated with other autoimmune diseases, including both hypothyroidism and Sjögren syndrome. Antimitochondrial antibodies, in particular, the anti-M2 isoform, are found with a high prevalence in patients with PBC. The anti-M2 antibodies are directed against the dihydrolipoamide acyltransferase component of the pyruvate decarboxylase complex.

Conversely, primary sclerosing cholangitis (PSC) is an inflammatory disease primarily affecting the extrahepatic bile ducts. Narrowing of the extrahepatic bile ducts results in clinical manifestations of right upper quadrant pain and jaundice; elevated ALP may be observed in individuals with PSC. PSC occurs predominantly in young to middle-aged males, and 80% also have ulcerative colitis. The primary autoantibodies associated with PSC, but not specific to the disease, are antineutrophil cytoplasmic antibodies. PSC disease progression may result in the development of cholangiocarcinoma, thereby worsening PSC-associated morbidity and mortality. Chronic cholestasis may result from mass lesions in the liver (tumors and granulomas as in sarcoidosis), with infiltrative processes in the liver (leukemia and amyloidosis), or with strictures of the bile ducts outside of the liver as well.

Hepatic tumors

HCC is the most common primary liver tumor. Worldwide, it is the fifth most common cause of cancer deaths. Anatomically, the liver is also a common site of metastases. The presence of hepatic tumors (primary or metastatic) may cause bile ductule obstruction, resulting in increased ALP and GGT activity; however, liver function tests may be normal. Chronic viral infections (HBV and HCV), cirrhosis, and NASH are all risk factors for the development of HCC. Approximately 70% of individuals with HCC also exhibit cirrhosis. Disease incidence is three-fold higher in males. Unfortunately, clinical manifestations do not present until late in the disease course. While there is interest in screening high-risk individuals, currently, the tumor marker AFP is used in monitoring disease progression, but not in diagnosis. Other markers such as the L3 variant of AFP and DCP (PIVKA-II) are more specific for HCC, but are not as sensitive as AFP, and are not widely used in the United States.

Suggested reading

[1] American Gastroenterological Association medical position statement: evaluation of liver chemistry tests. Gastroenterology. 2002; 123:1364.

[2] American Association for the Study of Liver Disease/Infectious Diseases, Hepatitis C guidance: AASLD-IDSA recommendations for testing, managing, and treating adults infected with hepatitis C, Hepatology. 62 (2015) 932–954.

[3] A.J. Czaja, Autoimmune liver disease, Curr. Opin. Gastroenterol. 25 (2009) 215–222.

[4] N. Chalasani, Z. Younossi, J.E. Lavine, et al., The diagnosis and management of non-alcoholic fatty liver disease: practice guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Society, Hepatology. 55 (2012) 2005–2023.

[5] D.R. Dufour, J.A. Lott, F.S. Nolte, D.R. Gretch, R.S. Koff, L.B. Seeff, Diagnosis and monitoring of hepatic injury. I. Performance characteristics of laboratory tests, Clin. Chem. 46 (2000) 2027–2049.

[6] D.R. Dufour, J.A. Lott, F.S. Nolte, D.R. Gretch, R.S. Koff, L.B. Seeff, Guidelines for use and interpretation of laboratory tests in hepatocellular injury. II. Recommendations for use of laboratory tests in diagnosis and monitoring of hepatic injury, Clin. Chem. 46 (2000) 2050–2068.

[7] D.R. Dufour, Alanine aminotransferase: is it healthy to be "normal"? Hepatology 50 (2009) 1699–1701.

[8] H.J. Kim, S.W. Oh, D.J. Kim, E.Y. Choi, Abundance of immunologically active alanine aminotransferase in sera of liver cirrhosis and hepatocellular carcinoma patients, Clin. Chem. 55 (5) (2009) 1022–1025.

[9] W. Lee, Acetaminophen-related acute liver failure in the United States, Hepatol. Res. 38 (2008) S3–S8.

[10] K.D. Lindor, M.E. Gershwin, R. Poupon, M. Kaplan, N.V. Bergasa, E.J. Heathcote, Primary biliary cirrhosis, Hepatology. 50 (2009) 291–308.

[11] J.M. Llovet, J. Bruix, Novel advancements in the management of hepatocellular carcinoma in 2008, J. Hepatol. 48 (Suppl 1) (2008) S20–S37.

[12] A.J. McCullough, R.S. O'Shea, S. Dasarathy, Diagnosis and management of alcoholic liver disease, J. Dig. Dis. 12 (2011). 257-26.

[13] Morbidity and Mortality Weekly Report (MMWR), Testing for HCV infection: an update of guidance for clinicians and laboratorians, MMWR. Cent. Dis. Control. Prev. 62 (2013) 362–365. 5.

[14] N.A. Terrault, N.H. Bzowej, K.M. Chang, J.P. Hwang, M.M. Jonas, M.H. Murad, AASLD guidelines for treatment of chronic hepatitis B, Hepatology. 63 (2016) 261–283.

Self-assessment questions

1. The synthetic functions of the liver are dependent on all of the following EXCEPT:
   a. Number of hepatocytes
   b. Blood flow from the portal vein
   c. Sufficient absorption of nutrients from the gastrointestinal tract
   d. Diameter of intrahepatic bile ducts
   e. Nutritional status
2. Which fraction of bilirubin is found in systemic circulation in a healthy individual?
   a. Conjugated bilirubin
   b. Unconjugated bilirubin
   c. Biliprotein
   d. Equal amounts of conjugated and unconjugated bilirubin
   e. Equal concentrations of biliprotein and unconjugated bilirubin
3. Which method for measuring total bilirubin concentrations uses caffeine as an accelerant?
   a. Evelyn–Malloy
   b. Jendrassik–Grof
   c. Van den Bergh and Muller
   d. Ehrlich
   e. An accelerant is not required for the measurement of total bilirubin concentrations.
4. Which disease is associated with increased unconjugated bilirubin concentrations at birth due to a complete loss of UDP-GT activity?
   a. Crigler–Najjar Type I
   b. Crigler–Najjar Type II
   c. Lucey–Driscoll
   d. Gilbert
   e. Dubin–Johnson
5. ALT requires which molecule as an enzymatic cofactor?
   a. Retinyl palmitate
   b. Zinc
   c. Magnesium
   d. Pyridoxal phosphate
   e. A cofactor is not required for enzymatic activity.
6. Which of the following antibodies is NOT associated with autoimmune hepatitis?
   a. Anti-LKM_1
   b. ASMA
   c. AMA-M2
   d. ANA
   e. All are associated with autoimmune hepatitis.
7. An increase of which of the following enzymes would be expected both in bone and liver diseases:
   a. Alanine aminotransferase
   b. Aspartate aminotransferase
   c. Alkaline phosphatase
   d. Gamma-glutamyl transferase
   e. Lactate dehydrogenase

For the following questions, use the following reference intervals: AST 5–40 IU/L; ALT 4–33 IU/L; alkaline phosphatase 40–130 IU/L; total protein 6.0–8.0 g/dL; albumin 3.7–5.0 g/dL; prothrombin time 12.0–14.3 s.

8. A patient is admitted to the hospital with the following laboratory results: AST 12,500, ALT 7300, alkaline phosphatase 125 (all IU/L), total and direct bilirubin 1.9 and 1.2 mg/dL, respectively, and prothrombin time of 22.0 s. The most likely cause of these abnormalities is:
   a. Alcoholic hepatitis
   b. Acute biliary tract obstruction
   c. Acute viral hepatitis
   d. Cirrhosis
   e. Toxic or ischemic hepatitis
9. A patient is admitted to the hospital with the following laboratory results: AST 280, ALT 95, alkaline phosphatase 85 (all IU/L), total and direct bilirubin 12.9 and 7.2 mg/dL, respectively, and prothrombin time of 12.0 s. The most likely cause of these abnormalities is:
   a. Alcoholic hepatitis
   b. Acute biliary tract obstruction
   c. Acute viral hepatitis
   d. Cirrhosis
   e. Toxic or ischemic hepatitis
10. A patient is admitted to the hospital with the following laboratory results: AST 140, ALT 150, alkaline phosphatase 380 (all IU/L), total and direct bilirubin 3.8 and 2.4 mg/dL, respectively, and prothrombin time of 12.8 s. The most likely cause of these abnormalities is:
    a. Alcoholic hepatitis
    b. Acute biliary tract obstruction
    c. Acute viral hepatitis
    d. Cirrhosis
    e. Toxic or ischemic hepatitis
11. All of the following findings in a person with chronic hepatitis would suggest progression to cirrhosis EXCEPT:
    a. AST/ALT ratio of >1
    b. Albumin of 4.0 g/dL
    c. Serum globulins of 6.5 g/dL
    d. Prothrombin time of 18.3 s
    e. Platelet count of 75,000/mm^3

Answers

1. d
2. b
3. b
4. a
5. d
6. c
7. c
8. e
9. a
10. b
11. b

Chapter 32

Clinical chemistry of the gastrointestinal disorders

David N. Alter

Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Describe the basic structure, function, and physiology of the GIT as it relates to digestion.
- Describe clinical laboratory tests used in the evaluation of gastritis.
- Describe clinical laboratory tests used in the evaluation of malabsorption.
- Describe clinical laboratory tests used in the assessment of GI neuroendocrine tumors.
- Describe clinical laboratory tests used in the evaluation of diarrhea and GI bleeding.

Discussion of gastrointestinal (GI) disease biomarkers requires a simultaneous review of GI tract (GIT) anatomy and physiology. The GIT represents a system of different organs similar in structure but with specific functions; thus markers of disease are specific to the targeted underlying pathology. Consequently, to organize the discussion along those terms, this review will proceed stepwisely through the GIT first in health and then in disease.

Anatomy/physiology

Digestion begins in the oral cavity where chewing facilitates digestion by increasing the surface area as well as stimulating salivary gland release of enzyme-rich saliva. Saliva contains salivary (S-type) amylase for the enzymatic breakdown of complex starches into simpler carbohydrates (e.g., maltose and dextrin) and provides lubrication for food transport through the esophagus (a skeletal and smooth muscle lined tube) into the stomach. The stomach has both mechanical and biochemical digestive functions and is geographically divided into the fundus, body, and antrum; each with different functions and associated cell types (Table 32.1). Relaxed, the stomach has about a volume of 75 mL but can distend to about 1 L when full.

Briefly, all gastric function is initially triggered by the vagal nerve-stimulated release of acetylcholine secondary to the sight, smell, and taste of food. This neural stimulation leads to a stepwise, but integrated, process of mechanical squeezing and distension of the stomach, secretion of histamine and gastrin (from G cells), and release of hydrochloric acid (HCl) (pH 1.5−3) and pepsinogen (I and II). As noted in Table 32.1, gastric cell types are distributed per function, with the protective (foveolar) and stimulatory (enteroendocrine) cells found relatively proximal, while "digestive" (parietal, chief) cells are more distal. Hyperacidity, coupled with the described mechanical processes, encompasses the gastric digestive process, namely, a physicochemical process designed to denature proteins for peptide chain release.

Clinically, gastroesophageal reflux disease [gastric contents reflux into the esophagus (normal pH 7.0)] and peptic ulcer disease (PUD; defects in gastric mucosa secondary to medications or disease exposing unprotected gastric mucosa to acidic conditions) are treated with H2 (histamine) blockers or proton pump inhibitors (PPIs). H2 blockers block histamine stimulation and subsequent parietal cell release of HCl. Conversely, PPIs shut down parietal cell ATPases required to transport actively ("pump") HCl against a millionfold concentration gradient into the gastric lumen. In both of those conditions, the goal is to reduce acid production as both disorders are associated with acid exposure to nongastric mucosa.

From the stomach, a bolus of semidigested food (chyme) enters the small intestine, where digestion continues and the majority of nutrient absorption occurs. The small intestine measures approximately 6−7 m in length by 3 cm in diameter with a surface area of 250 m^2 and is divided up into three geographic and functional segments:

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00032-6

TABLE 32.1 Gastric segments, cell types, and functions.

Segment	Cell type	Secretion	Function
Fundic body antral	Foveolar	Mucous	Mucosal layer for protection against gastric acid
		Pepsinogen II (antrum)	Protein breakdown
	Enteroendocrine	Gastrin and histamine	Stimulates antral motility Stimulates pepsinogen release from chief cells Stimulation of parietal cells to produce hydrochloric acid Stimulates secretion of GI hormones such as secretin, insulin, and ACTH
		Somatostatin	Inhibition of parietal cell gastric acid secretion opposing histamine and gastrin stimulation
Fundic	Parietal	Hydrochloric acid	Digestion
		Intrinsic factor	Vitamin B12 absorption
	Chief	Pepsinogen I	Protein breakdown

the duodenum (25 cm), jejunum (2.5 m), and ileum (3 m) with a gradually narrowing diameter from beginning to end. Digestion physically occurs in the GIT but is dependent on external organs such as the pancreas, liver, and gall bladder for pancreatic enzyme and proenzyme/bile synthesis, storage, and secretion. In the small intestine, chyme is broken down into its smallest components with enzymes (secreted locally and externally) specific for ingested carbohydrates, proteins, and fats. Briefly, carbohydrates are broken by amylase into monosaccharides (dextrose, fructose, and galactose) and oligosaccharides (dextrin and maltotriose). On the other hand, lipids and proteins undergo initial processing steps of emulsification and denaturation, respectively, to make them more digestible by their respective enzymes; lipase, colipase and bile salts; and gastric/pancreatic enzymes, for example, trypsinogen, chymotrypsinogen, and carboxypeptidases. Please refer to Chapter 33, Evaluation of exocrine pancreatic function, for more information on pancreatic enzyme activity and associated laboratory testing. Concomitant with food breakdown is nutrient absorption, which occurs over three phases: (1) luminal processing; (2) mucosal absorption; and (3) transport into circulation.

Duodenum

As chyme enters the small intestine; the pancreas secretes digestive enzymes and proenzymes for the biochemical digestion of digested material to produce amino acids, carbohydrates, and fats. In addition, bicarbonate is produced in pancreatic ductal cells to neutralize gastric acid and optimize enzymatic activity, along with the secretion of vasoactive intestinal peptide and secretin to inhibit gastrin and HCl production. In the duodenal mucosa, enterokinase is released and converts trypsinogen into trypsin. This activates a cascade leading to the conversion of pancreatic proenzymes into active molecules. The duodenum is also the primary site for iron absorption and can actively absorb calcium (when body stores are low).

Jejunum

The jejunum is the primary site of global (carbohydrate, lipid, and protein breakdown products) nutrient absorption (including passive calcium absorption). Absorption occurs at and is facilitated by numerous (20,000 in.^{2}) one to two cell layer thick fingerlike projections on the surface mucosa (villi) via active or passive transport mechanisms. Absorption of lipid breakdown products occurs mostly by passive diffusion or protein-mediated processes. In comparison, absorption of carbohydrate and amino acids or peptides occurs by active or passive transport mechanisms. With respect to amino acids and peptides, there are amino acid transporters as well as distinct di and tripeptide transporters to facilitate absorption.

Ileum

The ileum is similar in function to the jejunum; however, vitamin B12 (VB12) and bile acids are selectively absorbed in this region of the small intestine. Approximately, 80% of what is digested is subsequently absorbed by the small intestine, with an addition of 10% absorbed by the large intestine. The nonabsorbed residua enter the large intestine (colon; approximately 1−2-m long by 7 cm) for water and mineral reabsorption. Approximately, 8−9 L of fluid (oral—2 L, saliva—1.5 L, gastric—2.5 L, bile—0.5 L, and intestinal secretions—1.0 L) are absorbed daily by the gut (with a maximal absorption up to 20 L), and the remaining nonabsorbed material is excreted as feces.

Pathology

Practical biochemical markers for GIT-related disorders are limited in number. In addition, some markers discussed in the literature are either theoretical, minimally used, or historic. Consequently, the focus of this chapter will be on routinely utilized biochemical markers with mention of others as appropriate for comparison.

Oral cavity

There are numerous causes of oral pathology leading to impaired food intake; however, biochemical markers are of limited clinical utility, unless the pathology is associated with an underlying infectious etiology.

Esophagus

Most esophageal disorders are of either a functional (tears, ruptures, strictures, or congenital disorders), inflammatory (esophagitis), or malignant nature. Typically, esophageal disorders do not require routine laboratory-based biomarkers for diagnosis. However, iron deficiency, in combination with anemia, esophageal webs, cheilosis, and glossitis, is associated with *Paterson–Brown–Kelly* or *Plummer–Vinson syndrome*. Although the epidemiology/etiology of Plummer–Vinson syndrome is unclear, its overall incidence has declined. Of note, the Plummer–Vinson syndrome has been identified as a risk factor for the development of squamous cell carcinoma.

Stomach—gastritis

Gastritis (acute and chronic) is a common nonmalignant gastric inflammatory disorder in which there is a disruption of the gastric mucosal layer, allowing normal gastric acidic (pH 1.5–3) milieu to attack the underlying epithelium. The excess gastric acid can result in scarring, ulceration, and perforation. The clinical presentations of acute and chronic gastritis often overlap, and histologic examination is necessary for proper classification. For this discussion, nonatrophic chronic gastritis (superficial gastritis and diffuse antral gastritis) is of importance because of its association with acute *Helicobacter pylori* (*H. pylori*) infection and is typically limited to the antrum.

Clinically, gastritis can be divided into infectious, autoimmune, granulomatous, and unknown etiologies. *H. pylori* infection is the most common infectious cause of gastritis. Prominent noninfectious etiologies include pernicious anemia (PA), autoimmune thyroid disorders, gastric carcinoid tumors, and Crohn's disease (CD) [1]. Biochemically, gastrin and pepsinogen blood concentrations have been proposed in the scientific literature as markers of GI disorders; however, they have limited utility (discussed below).

Stomach—*Helicobacter pylori* infection

H. pylori infection is responsible for the majority of chronic gastritis and may be the most common chronic bacterial infection worldwide. *H. pylori* has elevated prevalence (>80%) in individuals greater than age 50 in resource-limited settings and developing countries. Fecal-oral and oral-oral routes are suspected transmission modalities for *H. pylori* acquisition. Indications for *H. pylori* testing vary, but several recommendations remain consistent across guidelines. Testing is indicated for the following scenarios: (1) low grade mucosal associated lymphoid tissue lymphoma; (2) active PUD or past history of peptic ulcer if cure of *H. pylori* infection has not been documented; and (3) early gastric cancer.

There is no gold standard method for the diagnosis of *H. pylori*; test selection may be influenced by predictive ability, local disease prevalence, clinical presentation, and anticipated course of action. For all nonserologic diagnostic tests, PPIs should be discontinued 1–2 weeks prior to testing and bismuth/antibiotic use should be discontinued up to 4 weeks prior. As will be discussed, urease produced by *H. pylori* is used in diagnosis by its ability to split urea into ammonia (NH_3) and carbon dioxide (CO_2).

Biopsy

Histologic identification of the organism and/or application of tissue to a urea-containing substrate (NH_3 production increases pH triggering a color change) may be performed to determine a *H. pylori* bacterial infection. However, endoscopy is not recommended for diagnosing a *H. pylori* bacterial infection, because it carries risks, has a high cost, and can result in false-negative results due to sampling error or missed organism detection.

Serologic testing

Several kits are available for the detection of *H. pylori* IgG antibodies. Overall, these assays have sensitivities approaching 100%, with specificities ranging from the mid-70s to mid-90s. They are rapid, inexpensive, and widely available for use; however, their appropriate utilization depends on local *H. pylori* prevalence; antibody testing is better suited for high-prevalence disease populations, presumably because of its high negative predictive value. Of note, serologic results can be positive for months to years posttreatment; thus serologic IgG testing does not distinguish between past and previous infections. *H. pylori* IgA and *H. pylori* IgM serologic tests are not widely used in clinical settings. In general, dependent on disease prevalence, diagnosis of *H. pylori* by serology is

rapidly fading due to the aforementioned inability to discriminate between past and current infection [2–5].

Breath/urea breath testing

An active *H. pylori* infection may be detected through measuring urea expiration from breath. From a testing perspective, a patient will ingest a capsule containing ^{13}C-labeled urea. Because *H. pylori* can convert urea into CO_2 and NH_3, radiolabeled $^{13}CO_2$ is produced and can be measured in breath. $^{13}CO_2$ is typically measured using spectrometric (such as infrared spectrometry) analysis. Urea breath testing has a specificity approaching 95% and sensitivity ranging from 88% to 95%. This approach is associated with a low probability of false-negative results, thereby allowing it to be used to document initial infection and treatment success. False-negative results are possible when the patient is being treated with antibiotics, PPIs, and bismuth. A 2–4 week period of nonuse is recommended prior to testing.

Stool testing

Detection of *H. pylori* antigen in stool is similar to the urea breath test in terms of test characteristics, with high sensitivities (>90%) and specificities (>90%), and has improved performance as compared with serologic IgG testing. *H. pylori* antigen testing has increased clinical specificity for a current infection. Thus many clinical laboratory environments have opted to use stool antigen testing to evaluate the presence of an active *H. pylori* infection. From a practical perspective, stool antigen testing is a kit-based ELISA test that is relatively easy to implement at the bench and is recommended as the initial test to use [5].

Stomach—pernicious anemia

PA is a specific form of VB12 deficiency associated with decreases in intrinsic factor (IF) production due to autoimmune destruction of parietal cells. PA is associated with chronic atrophic gastritis, IF autoantibodies, and gastric tumors. Once VB12 deficiency is confirmed via low VB12 and/or high methylmalonic acid (MMA) concentrations (low VB12 inhibits the conversion of methylmalonyl-CoA into succinyl-CoA causing an MMA accumulation), testing for anti-IF antibodies is recommended (sensitivity of 50%–70% with a specificity approaching 100%) [6].

Historically, the Schilling test was used to diagnose PA. However, this test has fallen out of favor, because it utilizes radioactive materials and is technically complicated to accurately perform. The Schilling test consists of successive stages of oral and intramuscular administration of radiolabeled VB12, with subsequent urine collection to determine if absorption is impaired; if results are abnormal, the test may be repeated with the addition of IF to determine if there is an IF-related defect.

Stomach—pepsinogen

Pepsinogen and the ratio of pepsinogen I and pepsinogen II have been discussed in the scientific literature as possible markers of atrophic gastritis and prognostic indicators of gastric cancer; however, metaanalyses of combined studies have revealed low sensitivities and specificities.

Stomach—gastrin

Hypergastrinemia is not a specific disease, but rather a finding with both gastric and nongastric causes. Gastrin elevations occur normally during the digestive process to increase gastric acidity with secretion inhibited by increased acidity (low pH). Gastrin secretion unopposed by gastric pH is considered pathologic. Causes of gastrin secretion relative to gastric pH are listed in Table 32.2. Gastrin elevations may be noted in multiple conditions; however, it is primarily employed as part of the workup for gastrinoma and Zollinger–Ellison syndrome. Both of these are discussed in more detail below.

Small intestine

Malignant tumors of the small intestine are not very common. They account for only 3% of GI malignancies (despite representing 90% of the total GI surface area) and 0.5% of cancers overall. However, in terms of biochemical monitoring of GI disorders, a discussion of two relatively common small intestine related neoplasms, gastrinoma and carcinoid, is required.

TABLE 32.2 Gastrin elevation.

Appropriate—decreased acid production (neutral pH)
Food
Histamine blocker/PPI therapy
H. pylori
Atrophic gastritis
Vagotomy
Inappropriate—increased acid production (low pH)
Gastrinoma/Zollinger–Ellison syndrome
Retained antrum syndrome
Antral G cell hyperplasia
Pyloric (gastric outlet) obstruction
Renal failure

PPI, Proton pump inhibitor.

Small intestine—duodenum—gastrinoma

A gastrinoma is a well-differentiated neuroendocrine tumor (NET) of the duodenum (15%) or pancreas (85%) that causes hypersecretion of gastrin and can lead to Zollinger—Ellison syndrome (gastric acid excess, gastric acid-related gastritis, and diarrhea) [6]. Clinically, gastrinomas are difficult to diagnose due to their nonspecific clinical presentation (abdominal pain and chronic diarrhea). Diagnostic for gastrinoma is the appropriate clinical presentation and a serum gastrin concentration greater than 10 × the upper limit of normal (typically >1000 pg/mL) in the presence of gastric acid pH < 4. Secretin stimulation and calcium infusion dynamic testing may also be used to distinguish gastrinoma from other causes of hypergastrinemia; however, the former is preferred over the latter.

Small intestine—carcinoid

Carcinoid tumors are a general class of well-differentiated NETs with a very low overall incidence (carcinoid tumor: 38/1,000,000) but with characteristic clinical and biochemical presentation (carcinoid syndrome, described below). Carcinoid tumors are predominantly located in the GIT or pulmonary system (55%—30%, respectively). Within the GIT, carcinoid tumors are referred to as NETs and are classified/stratified in terms of clinical characteristics and prognosis by embryonic derivation; foregut (intrathoracic, gastric, and duodenal); midgut (small intestine, appendix, and proximal colon), and hindgut (distal colon and rectum). A description of GIT carcinoid characteristics is described in Table 32.3. In general, these tumors clinically present via their production of vasoactive compounds ("carcinoid syndrome"). Pancreatic carcinoid is very rare and needs to be distinguished from pancreatic endocrine tumors and their respective specific syndromes (e.g., insulinomas, gastrinomas, VIPomas glucagonomas, and somatostatinomas). Prognostically, carcinoid metastatic potential is directly related to size; however, colonic, well-differentiated NETs are more aggressive than any other area of the GIT [7,8].

Small intestine—carcinoid syndrome

Biochemically, carcinoid tumors are characterized by excess serotonin production and release, resulting in the clinical findings associated with the carcinoid syndrome, including flushing, diarrhea, bronchoconstriction, and cardiac valvular lesions (Fig. 32.1). Diagnosis is based on the detection of the serotonin metabolite [5-hydroxyindolacetic acid (5-HIAA)] in a 24-hour urine specimen (sensitivity >90% and specificity 90%). Tryptophan- and serotonin-rich foods, such as banana, pineapple, tomato, plums, eggplant, avocado, kiwi, and nuts, should be

TABLE 32.3 Gastrointestinal carcinoid distribution and characteristics.

Location	GIT incidence	Comments
Stomach (foregut)	7%—9%	• Often incidental identification but can be symptomatic (10% with carcinoid syndrome; pain, anemia, and GI bleeding). • Type 1, 70%—80% with atrophic gastritis and achlorhydria • Type 2, 5% associated with gastrinoma • Type 3, 20% sporadic and the most aggressive are associated with elevated gastrin concentrations • Low DOPA decarboxylase activity, and therefore low 5-HIAA and serotonin levels. Diagnosis is based on 5-HTP levels.
Small intestine (midgut)	38%—45%	• 90% in distal ileum • >60% are metastatic at presentation (regional + distant) • Clinical presentation as carcinoid syndrome (up to 13% of cases) secondary to serotonin release (see discussion). Diagnosis is based on urinary 5-HIAA detection.
Colon (midgut and hindgut)	12%	Lack DOPA decarboxylase, and therefore unable to convert 5-HT into serotonin.
Appendix	16%—18%	Rarely symptomatic and usually incidental detection.
Rectum	20%	Incidental identification can present with rectal bleeding and pain.
Pancreas	Rare	• Large (up to 7 cm) with frequent presentation of carcinoid syndrome (up to 100%) and metastatic spread. • Pancreatic serotonin producing tumors (carcinoid) are an extremely rare subset of pancreatic neuroendocrine tumors.

5-HIAA, 5-hydroxyindolacetic acid; *5-HT*, 5-hydroxytryptophan; *GIT*, gastrointestinal tract.

Tryptophan
↓ *Tryptophan-5-hydroxylase*
5-Hydroxytrytophan (5-HTP)
↓ *Aromatic L-amino acid decarboxylase*
5-Hydroxytryptamine (5-HT; Serotonin)
↓ *Monoamine oxidase*
5-Hydroxyindoleacetaldehyde
↓ *Aldehyde dehydrogenase*
5-Hydroxyindoleacetic acid (5-hiaa) *Excreted in urine* *Utilized for diagnosis (24-hour urine specimen).*

FIGURE 32.1 Biochemistry of carcinoid syndrome and diagnosis.

avoided up to 24 hours prior to beginning of urine collection to prevent false-positive results. 5-HIAA is most useful for patients with midgut tumors.

Foregut and hindgut-related carcinoid tumors have low to no activity of the rate limiting enzyme [aromatic L-amino acid decarboxylase (DOPA decarboxylase)] responsible for serotonin production; therefore, in these types of tumors, 5-HIAA testing has limited utility. Chromogranin A, a protein stored within neuroendocrine tissues, has been discussed in the literature as an additional biochemical marker of carcinoid tumors, but it has poor test characteristics to be a practical diagnostic tool. A better use of chromogranin A may be as a marker of carcinoid treatment and prognosis.

Small intestine—malabsorption

Clinically, malabsorption symptoms can range from asymptomatic through nonspecific symptoms (anorexia, flatulence, and abdominal distension) to, in severe cases, weight loss, diarrhea, and even death. Malabsorption has both small intestinal and nonsmall intestinal causes; however, in all cases, there is an inability to absorb nutrients. For example, PA and pancreatic enzyme deficiencies are extraintestinal, but result in intestinal malabsorption syndromes.

Small intestine—fat malabsorption

The most common cause of fat malabsorption is due to pancreatic insufficiency or pancreatic enzyme inactivation. Clinically, patients commonly present with steatorrhea; fat malabsorption can be assessed by fecal fat analysis (discussed later in this chapter).

Small intestine—carbohydrate malabsorption

Carbohydrate digestion occurs throughout the GIT starting in the mouth (S-type amylase) and continues utilizing pancreatic (P-type) amylase followed by enzymes located on the "brush border" of the small intestinal surface epithelium (sucrase, lactase, and isomaltase). Enzyme deficiencies from the pancreas, or disorders that reduce functional intestinal epithelium availability at the brush border, may result in decreased serum glucose concentrations as well as the buildup of undigested carbohydrates; the latter can then ferment producing hydrogen gas. Diagnostically, serum concentrations of the target carbohydrate, as well as their subsequent fermentation product, are utilized in assessment of carbohydrate malabsorption.

D-Xylose test

Administration of D-xylose followed by serum or urine measurements of D-xylose is performed to measure proximal small intestine absorptive (not digestive) capacity. If the serum or urine concentrations of D-xylose are lower than expected (indicating that the molecule was not absorbed), a mucosal absorptive defect is suspected. This test is subject to false positives, and may be elevated in patients with renal disease.

Lactose tolerance test

Administration of a measured lactose dose is followed by blood glucose measurements. Glucose is measured as lactose is cleaved by lactase into glucose and galactose and galactose is subsequently converted into glucose in the liver. The absence of elevations in glucose concentrations, in conjunction with clinical symptoms, is diagnostic of carbohydrate malabsorption. However, false-negative results may occur in cases of diabetes (due abnormal carbohydrate metabolism) or in the background of intestinal bacterial overgrowth.

Breath testing

Administration of a measured carbohydrate dose (lactose, fructose, and sucrose) is followed by measurement of hydrogen (H_2) breath levels over a 3-hour period. The pattern of H_2 breath levels at distinct times may be used to evaluate lactose tolerance. A certain pattern of these results is diagnostic of small intestinal bacterial overgrowth (SIBO). Methane testing is also recommended as a small percentage of cases will have nonhydrogen producing bacteria [9–11].

TABLE 32.4 Conditions associated with small intestinal bacterial overgrowth.

Celiac disease
Short bowel syndrome
Chronic pancreatitis
Intestinal fistula
Immunodeficiency
Liver disease
Diabetes mellitus
Scleroderma
Radiation enteritis
Inflammatory bowel disease
Irritable bowel syndrome
Cystic fibrosis

Small intestine—celiac disease

Celiac disease is an autoimmune-mediated intestinal epithelium inflammatory disorder that impairs nutrient absorption. This immune-mediated disorder affects about 1% of the population and is triggered by prolamins (plant-based storage protein) found in wheat (gliadin) and other grains such as barley and rye. Clinically, it manifests as chronic diarrhea, flatulence, abdominal distension/pain, and weight loss. Treatment involves cessation of these foods and is very successful. Testing should be performed after the individual has been on a gluten containing diet as false-positive results can arise, if done otherwise. Serologic testing for antibodies to tissue transglutaminase (TTG; IgA-TTG) has proven to be the optimal serum-based laboratory test in the assessment of celiac disease. Prior to the use and development of TTG serology, antiendomysial antibodies, and antigliadin (IgA and IgG) antibodies had been commonly utilized; however, their test characteristics were not as sensitive or specific as TTG. Low-risk patients with positive serologic testing and high-risk patients regardless of serology should get biopsied. With respect to celiac disease (as well as many others), laboratories should work in concert with local allergists to develop the best protocol for their specific patient population [12,13].

Small intestine—bacterial overgrowth

SIBO syndrome is characterized by an overgrowth of native and/or nonnative bacteria due to poorly absorbed nutrients that result in excessive fermentation and inflammation. Bacterial overgrowth can decrease nutrient absorption, thereby providing a conducive environment for bacterial growth and further exacerbating the underlying condition. Associated disorders are summarized in Table 32.4. Nonspecific clinical and laboratory findings associated with SIBO include abdominal discomfort, bloating, diarrhea, weight loss, macrocytic anemia (VB12 deficiency), increased fecal fat, hypoalbuminemia, and protein deficiency. Confirmation of SIBO requires a combination of laboratory tests, clinical findings, endoscopy, and knowledge of any underlying medical disorders. Breath testing is discussed in the literature and has been occasionally used to help make the diagnosis of SIBO. Breath testing is based on the principle that the overgrown bacteria will cause increased production and expiration of hydrogen gas ($\geq$20 ppm by 90 minutes) or methane [9–11].

Colon

Significant disorders of the large intestine usually present clinically as either diarrhea, bleeding (gross to occult presentation), or a combination of both.

Colon—diarrhea

Definitions vary (i.e., increases in stool by weight, volume, or frequency) but consistent features are persistent liquidity with poorly formed stools. In general, it represents an inability of the large intestine to reabsorb excess water and water-soluble electrolytes passing through the GIT. Diarrhea is broadly classified into two categories: osmotic and secretory; additional causes include inflammatory (colitis-related) and altered motility (vagotomy, sympathectomy, adrenal insufficiency, hyperthyroidism, and irritable bowel syndrome).

Osmotic causes are due to poorly or nonabsorbed substances (e.g., undigestable sugar alcohols, e.g., sorbitol and mannitol) that prevent normal water reabsorption by shifting the osmotic gradient to favor water retention over reabsorption. Osmotic diarrhea should resolve with cessation of the offending agent. In contrast, secretory diarrhea has a multitude of etiologies that include, but are not limited to, bacterial toxins inflammatory bowel disease, lymphocytic colitis and medications.

Diarrheal workup should include routine laboratory testing for blood count, electrolyte, and hepatic and renal function tests to assess patient status and because the differential diagnosis is protean in nature [14,15].

Stool electrolytes

Stool electrolytes are not osmotically active (similar to plasma) and are used to calculate a stool osmotic gap based on the difference between a measured and calculated stool osmolality.

$$\text{Stool Osm gap} = \left(290\left(\text{stool osmolality mOsm/Kg}\right)\right) - 2(\text{Na} + \text{K})\ \text{mOsm/Kg}$$

$<$50 mOsm/Kg) Consistent with secretory type diarrhea

$>$100 mOsm/Kg) Consistent with osmotic type diarrhea

Stool pH values less than 6 are consistent with fermentation and are suggestive of carbohydrate malabsorption. Stool submitted for diarrheal workup, regardless of test, is only useful if liquid stool is submitted. Lastly, stool may be submitted for fecal fat analysis to diagnose steatorrhea. According to the literature and labs that still perform the test, a 24−72-hour (72 preferred) specimen with a specific fat-controlled (100−150-g fat/day) diet during testing is required. In addition, 3 days prior and during the testing period, laxatives and synthetic fat substitutes should be avoided. Expected values are <7-g fat/24 hours (timed specimen) or 0%−19% fat (random collection). A recent discussion has proposed an alternative strategy that depends on stool weight and detection of fat globules [16]. Due to variability in collection and patient preparation, several test utilization campaigns do not endorse using 72-hour quantitative fecal fat testing in the evaluation of lipid malabsorption.

Colon—*Clostridium difficile*-related diarrhea

Any discussion about diarrheal workups would be incomplete without mention of *Clostridium difficile*-related diarrhea and its associated evaluation. *C. difficile* is a Gram-positive spore-forming anaerobic bacterium found in soil and GIT that produces virulent exotoxins (toxin A and toxin B) responsible for colitis and diarrhea. For *C. difficile* infection (CDI) to occur, both the organism and disruption of normal GIT microbial flora are required. As normal flora gets disrupted under conditions, such as, but not limited to, increased age, prolonged hospitalization, prolonged antibiotic use, PPIs, chemotherapy, and chronic kidney disease, conditions become favorable for existing *C. difficile* spores to become "active" by overwhelming normal host defenses releasing their toxins. These interfere with intestinal epithelial integrity, causing an inflammatory mediator release and fluid shift, which lead to inflammation and diarrhea. The prevalence of asymptomatic carriage of *C. difficile* spores is estimated at about 2% in the healthy population, with 30% prevalence in inpatients. However, frequency is dependent on underlying morbidity, antibiotic use, and length of hospitalization.

Diagnosis of *C. difficile* includes diarrhea (three documented episodes per day) in the setting of recent or ongoing antibiotic treatment, especially in hospitalized and/or older individuals. Presumably, episodes of diarrhea in prolonged care (hospital, long-term care facility, nursing home, etc.) should also be investigated for the possibility of *C. difficile* due to contributing comorbidities. Diagnostic modalities are based on either detection of the organism or detection of the toxin. Relevant testing is summarized in Table 32.5. Regardless of method used, the following general principles must be considered: (1) diarrhea must be confirmed; (2) the patient must be symptomatic; and (3) the specimen must be a liquid stool. Assessment of stool liquidity, that is, appropriateness for *C. difficile* evaluation, is accomplished using either the Bristol Stool Form Scale (1−7 with 7 = "liquid or taking the form of a cup") or the Brecher criteria ("if the stick falls, test them all. If the stick stands, the test is banned"). In addition, testing cannot be used to determine "cure," because treated/"cured" (asymptomatic) individuals can continue to shed the organism for up to 6 weeks. Finally, CDI severity stratification includes the use of white blood cell (WBC) count and serum creatinine with increasing values of both (>15,000 and >1.5× baseline, respectively), which are consistent with worsening severity [17−19].

Colon—gastrointestinal bleeding

The differential diagnosis of GIT bleeding is long and can be due to multiple unrelated causes ranging from gingival disease and long distance running through any disorder discussed in this chapter to colonic malignancy. Location/rate/volume of bleed determines how blood appears in stool with presentations varying from gross blood to melena (dark/tarry stools) to not seen at all (occult). If gross blood is observed, identification of blood is usually obvious. However, melena can be a subjective interpretation as charcoal and diet can result in stools with a similar appearance requiring additional testing. Occult refers to "hidden" and fecal occult blood (FOB) indicates blood in the stool that cannot be seen. Evaluation for FOB is performed to assess for unexplained anemia, as a screening test for cancer and in the presence of melena stools. In terms of colon malignancy, occult hemorrhage of the colonic mucosa occurs and increases as precancerous lesions progress to cancerous secondary to an increasing vascularity, cellularity, and impact to the colonic mucosa.

Fecal occult blood testing

Fecal occult blood testing (FOBT) targets either heme or globin components of hemoglobin. Both methods have their pros and cons but globin is gradually replacing the heme method. Detection of the heme component has been around since the 1950s utilizing the ubiquitous plant extract guaiac. Application of the peroxide "developer" in the presence of heme (a peroxidase) will release oxygen, subsequently reacting with alpha-guaiaconic acid impregnated on the card (guiac is extracted from the resin of the hardwood tree *Guaiacum officinale)* to produce a blue-colored product. Operationally, this test is performed by application of stool samples to guaiac pretreated slides followed by two drops of "developer" (hydrogen peroxide solution) and then visual examination of the slide for a

TABLE 32.5 ***Clostridium difficile*** **diagnosis.**

Toxigenic culture	• Reference standard • Isolation of organism from feces followed by confirmation that isolate is toxin producing strain • Time-consuming (24–48 h) and labor-intensive • "Very sensitive"
Cell cytotoxicity assay	• Reference standard • Cell culture-based assay that identifies cytopathic effect of related toxins with confirmatory step to confirm *C. difficile* toxins • Time-consuming (24–48 h) and labor-intensive • Sensitivities 65%—90%
There is some debate in the reviewed literature as to which test TC versus CCA are better reference methods for the detection of *C. difficile*.	
Immunoasssay—GDH	• Metabolic enzyme produced in all strains (toxigenic and nontoxigenic) of *C. difficile* • Up to 46% *C. difficile* strains have been reported to be nontoxigenic • Suitable as a screening test but positive results require confirmation testing • Sensitivities and specificities in the high 1990s • Low-cost kit-based EIA with rapid turnaround time
Immunoassay—toxins (A and B)	• Monoclonal/polyclonal antibodies for detection of toxins A and B • Toxin A can be nontoxigenic, so A and B must be included • Sensitivities ranging from 40% to 99% with specificities ranging 60% to 100% • Low-cost kit-based EIA with rapid turnaround time
There are some kits available on the market, which combine GDH and toxin detection with sensitivities in the 50%–70% range and specificities approaching 100%. However, specimens that are GDH positive and toxin negative should have a confirmatory test performed.	
Molecular—NAAT/PCR	• Molecular detection of either *C. difficile* toxin-related gene sequences to variety of specific methods • Real-time automated molecular technology • Test characteristics are dependent on method/manufacturer utilized with sensitivities ranging from 80% to 100% and specificities approaching 100%

GDH, Glutamate dehydrogenase.

blue color change. Preanalytically, it is important to not apply stool thickly and to confirm that the control circles have properly reacted. The test (FOBT) is subject to dietary false-positive results (upper GI bleeding, animal heme, and vegetable peroxidases) and to false-negative results [antioxidant ingestion such as ascorbic acid (vitamin C)].

Slowly replacing the FOB is the fecal immunochemical test (FIT), which utilizes antibodies specific for the intact hemoglobin molecule. The FIT does not suffer from the false positives and false negatives of the FOBT, but has some limitations itself. Degradation of the globin chain from time/temperature factors can occur (the heme moiety does not) and the test's ability has not been sufficiently evaluated for patients with globin chain disorders. In addition, as the globin chain will be digested, the FIT will not detect the presence of upper GI bleeding. Test characteristics for FOBT and FIT are compared and contrasted in Table 32.6.

No discussion of FIT and FOBT would be complete without mention of stool-based molecular testing being used for colorectal cancer screen. Currently, there is a commercial (Cologuard) test available that assays stool for KRAS mutations, methylation markers, and hemoglobin. For this method, patients collect their stool and send it directly to the company, where it is then analyzed with a proprietary algorithm and methodology. FOBT and FIT have shown that they do detect polyp early and have reduced colorectal cancer mortality. With respect to the molecular test, there appears to be a benefit but data are still limited.

Colon—tumor markers

In terms of colon cancer, carcinoembryonic acid (CEA) is the biochemical (numerous molecular markers are available) tumor marker used in practice. CEA is a glycoprotein associated with cell adhesion produced during fetal development. As well as colorectal cancer, CEA can be elevated in other conditions such as, but not limited to, tobacco smokers, gastritis, inflammatory bowel disease (IBD), liver disease, and chronic obstructive pulmonary disease. The test has a poor sensitivity (46%) and specificity (89%) for diagnosis and is not recommended for screening or diagnosis. However, it may be used to monitor treatment and cancer recurrence [20].

TABLE 32.6 Test characteristics of fecal occult blood testing and fecal immunochemical test.

	FOBT	FIT
Advantages	Point-of-care test that is easy to perform and cards can be sent home for patient application of stool Heme not subject to digestion environmental degradation	Not subject to dietary interference Serial testing not required
Disadvantages	False-positive and false-negative dietary impacts Sensitivity improves with serial testing	Laboratory-based test requiring some manipulation False-negative results possible due to digestion or environmental degradation Will not detect presence of upper GI bleeding
Sensitivity	Single test 24%–40% Serial [3] test 80%–92% Varies by manufacturer and formulation	64%–80% Varies by manufacturer
Specificity		87%–98% Varies by manufacturer

FIT, Fecal immunochemical test; *FOBT*, fecal occult blood testing; *GI*, gastrointestinal.

Colon—inflammatory bowel disease

IBD is delineated into two major types: ulcerative colitis (UC) and CD. UC is characterized by mucosal limited inflammation restricted to the colon, whereas CD appears throughout the GIT and has a transmural inflammatory response. Both diseases clinically present with diarrhea and abdominal pain with diagnosis based on history, clinical presentation, and nonspecific markers of systemic inflammation such as WBC count or C-reactive protein elevation. There are also biochemical markers specific to colonic inflammation, namely, fecal leukocytes, lactoferrin, and calprotectin. All these markers have varying degrees of evidence-based support and have shown greatest utility in ruling out IBD; these tests also have a reported utility for treatment and relapse monitoring.

Fecal leukocyte (WBC) testing has been in use since the 1970s as a tool to diagnose infectious diarrheas especially salmonellosis and shigellosis. The test depends on the infection's ability to spill WBCs into the intestinal lumen; therefore, its diagnostic performance depends on type and severity of infection as well as the ability of the operator to microscopically identify WBCs in stool specimens. Test specificity has been reported to be in the 1970s, with sensitivities ranging from 30% to 80% depending on study reviewed. Better than identification of WBCs has been the development of immunoassays for WBC (neutrophil)-associated constituents, namely, lactoferrin (iron binding protein from neutrophil secondary granules) and calprotectin (calcium–zinc binding protein neutrophilic cytosolic protein). For the detection of inflammatory disorders especially IBD, both have reported sensitivities in the 1990s with specificities approaching 100%. Neither can be used to distinguish UC from CD, but they can monitor treatment and relapse as needed. Unfortunately, both will be also be present in most intestinal inflammatory disorders. In addition, neither is useful in distinguishing *C. difficile*-related diarrhea from other etiologies. To differentiate biochemically between CD and US, two serologic markers are currently being used: antisaccaromyces cervisiae antibodies (ASCA) and perinuclear antineutrophil cytoplasmic antibodies (pANCA). In short, various combinations of ASCA and pANCA positivity and negativity have sensitivities in the 1950s to 1960s and specificities in the high 1980s and 1990s [21–23].

References

[1] N.A. Abumrad, F. Nassir, A. Marcus, Digestion and absorption of dietary fat, carbohydrate and protein, in: M. Feldman, L.S. Friedman, L.J. Brandt (Eds.), Sleisenger and Fordtran's Gastrointestinal Liver Disease, Saunders, Philadelphia, PA, 2015, pp. 1736–1764.

[2] M.C. Rao, J. Sarathy, J.H. Sellin, Intestinal electrolyte absorption and secretion, in: M. Feldman, L.S. Friedman, L.J. Brandt (Eds.), Sleisenger and Fordtran's Gastrointestinal Liver Disease, Saunders, Philadelphia, PA, 2015, pp. 1713–1736.

[3] C.C. Robayo-Torres, R. Quezada-Calvillo, B.L. Nichols, Disaccharide digestion: clinical and molecular aspects, Clin. Gastroenterol. Hepatol. 4 (3) (2006) 276–287.

[4] N.B. Vakil, Physiology of gastric acid secretion, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA, 2018 (accessed 11.01.18).

[5] J.R. Turner, The gastrointestinal tract, in: V. Kumar, A.K. Abbas, J.C. Aster (Eds.), Robbins and Cotran Pathologic Basis of Disease, Saunders, Philadelphia, PA, 2014, pp. 749–849.

[6] M. Feldman, P.J. Jensen, Classification and diagnosis of gastritis and gastropathy, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA, 2019 (accessed 11.01.18).

[7] M. Feldman, E.L. Lee, Gastritis, in: M. Feldman, L.S. Friedman, L.J. Brandt (Eds.), Sleisenger and Fordtran's Gastrointestinal Liver Disease, Saunders, Philadelphia, PA, 2015, pp. 868–883.
[8] M. Varbanova, K. Frauenschläger, P. Malfertheiner, Chronic gastritis - an update, Best. Pract. Res. Clin. Gastroenterol. 28 (6) (2014) 1031–1042.
[9] W.D. Chey, G.I. Leontiadis, C.W. Howden, S.F. Moss, ACG Clinical Guideline: treatment of *Helicobacter pylori* infection, Am. J. Gastroenterol. 112 (2017) 212–239.
[10] S.E. Crowe, Indications and diagnostic tests for *Helicobacter pylori* infection, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA, 2019 (accessed 11.01.18).
[11] P. Sabbagh, M. Mohammadnia-Afrouzi, M. Javanian, A. Babazadeh, V. Koppolu, V.R. Vasigala, et al., Diagnostic methods for *Helicobacter pylori* infection: ideals, options, and limitations, Eur. J. Clin. Microbiol. Infect. Dis. (2018).
[12] D.R. Morgan, S.E. Crowe, *Helicobacter pylori* infection, in: M. Feldman, L.S. Friedman, L.J. Brandt (Eds.), Sleisenger and Fordtran's Gastrointestinal Liver Disease, Saunders, Philadelphia, PA, 2015, pp. 856–867.
[13] S.K. Patel, C.B. Pratap, A.K. Jain, A.K. Gulati, G. Nath, Diagnosis of *Helicobacter pylori*: what should be the gold standard? World J. Gastroenterol. 20 (36) (2014) 12847–12859.
[14] R.C. She, A.R. Wilson, C.M. Litwin, Evaluation of *Helicobacter pylori* immunoglobulin G (IgG), IgA, and IgM serologic testing compared to stool antigen testing, Clin. Vaccine Immunol. 16 (8) (2009) 1253–1255.
[15] C.M. Rojas Hernandez, T.H. Oo, Advances in mechanisms, diagnosis, and treatment of pernicious anemia, Discov. Med. 19 (104) (2015) 159–168.
[16] Y. Kitamura, M. Yoshihara, M. Ito, T. Boda, T. Matsuo, T. Kotachi, et al., Diagnosis of *Helicobacter pylori*-induced gastritis by serum pepsinogen levels, J. Gastroenterol. Hepatol. 30 (10) (2015) 1473–1477.
[17] R. Arnold, Diagnosis and differential diagnosis of hypergastrinemia, Wien. Klin. Wochenschr. 119 (19–20) (2007) 564–569.
[18] S. Dacha, M. Razvi, J. Massaad, Q. Cai, M. Wehbi, Hypergastrinemia, Gastroenterol. Rep. (Oxf.) 3 (3) (2015) 201–208.
[19] J.F. Rehfeld, L. Bardram, L. Hilsted, P. Poitras, J.P. Goetze, Pitfalls in diagnostic gastrin measurements, Clin. Chem. 58 (5) (2012) 831–836.
[20] I. Epelboym, H. Mazeh, Zollinger-Ellison syndrome: classical considerations and current controversies, Oncologist. 19 (1) (2014) 44–50.
[21] R.T. Jensen, J.A. Norton, K. Oberg, Neuroendocrine tumors, in: M. Feldman, L.S. Friedman, L.J. Brandt (Eds.), Sleisenger and Fordtran's Gastrointestinal Liver Disease, Saunders, Philadelphia, PA, 2015, pp. 501–541.
[22] M.A. Maggard, J.B. O'Connell, C.Y. Ko, Updated population-based review of carcinoid tumors, Ann. Surg. 240 (1) (2004) 117–122.
[23] K. Oberg, Neuroendocrine gastrointestinal and lung tumors, the carcinoid syndrome and related disorders, in: S. Melmed, K.S. Polonsky, P.R. Larsen, H.M. Kronenberg (Eds.), Williams Textbook of Endocrinology, Saunders, Philadelphia, PA, 2013, pp. 1833–1853.

Further reading

B.P. Abraham, S. Kane, Fecal markers: calprotectin and lactoferrin, Gastroenterol. Clin. North. Am. 41 (2) (2012) 483–495.
N. Bagdasarian, K. Rao, P.N. Malani, Diagnosis and treatment of *Clostridium difficile* in adults: a systematic review, JAMA. 313 (4) (2015) 398–408.
V.E. Barakauskas, G.Y. Lam, M.P. Estey, Digesting all the options: laboratory testing for celiac disease, Crit. Rev. Clin. Lab. Sci. 51 (6) (2014) 358–378.
D. Basso, C.F. Zambon, M. Plebani, Inflammatory bowel diseases: from pathogenesis to laboratory testing, Clin. Chem. Lab. Med. 52 (4) (2014) 471–481.
I. Brusca, Overview of biomarkers for diagnosis and monitoring of celiac disease, Adv. Clin. Chem. 68 (2015) 1–55.
C.A. Burnham, K.C. Carroll, Diagnosis of Clostridium difficile infection: an ongoing conundrum for clinicians and for clinical laboratories, Clin. Microbiol. Rev. 26 (3) (2013) 604–630.
M.R. Carroll, H.E. Seaman, S.P. Halloran, Tests and investigations for colorectal cancer screening, Clin. Biochem. 47 (10–11) (2014) 921–939.
A. Duncan, C. Robertson, R.I. Russell, The fecal osmotic gap: technical aspects regarding its calculation, J. Lab. Clin. Med. 119 (4) (1992) 359–363.
P.H. Gilligan, Optimizing the laboratory diagnosis of Clostridium difficile infection, Clin. Lab. Med. 35 (2) (2015) 299–312.
C. Hogenauer, H.F. Hammer, Maldigestion and malabsorption, in: M. Feldman, L.S. Friedman, L.J. Brandt (Eds.), Sleisenger and Fordtran's Gastrointestinal Liver Disease, Saunders, Philadelphia, PA, 2015, pp. 1788–1823.
C.P. Kelly, Diagnosis of celiac disease in adults, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA, 2019 (accessed 11.01.18).
M. Korostensky, S.R. Martin, M. Swain, M. Raman, T. Naugler, S.M. Hossein-Sadrzadeh, et al., Elimination of 72-hour quantitative fecal fat testing by restriction, laboratory consultation, and evaluation of specimen weight and fat globules, J. Appl. Lab. Med. 3 (2018) 357–365.
J.B. Mason, V. Milovic, Clinical features and diagnosis of malabsorption, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA, 2019 (accessed 11.01.18).
P. Moayyedi, ACP Journal Club. Review: calprotectin testing differentiates inflammatory bowel disease from the irritable bowel syndrome, Ann. Intern. Med. 160 (8) (2014).
J.D. Ostrow, Tests for fecal occult blood, in: H.K. Walker, W.D. Hall, J. W. Hurst (Eds.), Clinical Methods: The History, Physical, and Laboratory Examinations, Butterworths, Boston, MA, 1990. <http://www.ncbi.nlm.nih.gov/books/NBK445/> (accessed 15.12.15).
T. Planche, M.H. Wilcox, Diagnostic pitfalls in Clostridium difficile infection, Infect. Dis. Clin. North. Am. 29 (1) (2015) 63–82.
E.M. Quigley, Small intestinal bacterial overgrowth, in: M. Feldman, L. S. Friedman, L.J. Brandt (Eds.), Sleisenger and Fordtran's Gastrointestinal Liver Disease, Saunders, Philadelphia, PA, 2015, pp. 1824–1831.
A. Rezaie, M. Buresi, A. Lembo, H. Lin, R. McCallum, S. Rao, et al., Hydrogen and methane-based breath testing in gastrointestinal disorders: the North American Consensus, Am. J. Gastroenterol. 112 (2017) 775–784.
R.J. Saad, W.D. Chey, Breath testing for small intestinal bacterial overgrowth: maximizing test accuracy, Clin. Gastroenterol. Hepatol. 12 (12) (2014) 1964–1972.

B.E. Sands, Biomarkers of inflammation in inflammatory bowel disease, Gastroenterology. 149 (5) (2015) 1275–1285.
T.J. Savides, D.M. Jensen, Diarrhea, in: M. Feldman, L.S. Friedman, L.J. Brandt (Eds.), Sleisenger and Fordtran's Gastrointestinal Liver Disease, Saunders, Philadelphia, PA, 2015, pp. 221–241.
T.J. Savides, D.M. Jensen, Gastrointestinal bleeding, in: M. Feldman, L. S. Friedman, L.J. Brandt (Eds.), Sleisenger and Fordtran's Gastrointestinal Liver Disease, Saunders, Philiadelphia, PA, 2015, pp. 297–335.
L.R. Schiller, J.R. Sellin, Diarrhea, in: M. Feldman, L.S. Friedman, L.J. Brandt (Eds.), Sleisenger and Fordtran's Gastrointestinal Liver Disease, Saunders, Philadelphia, PA, 2015, pp. 221–242.
Y.F. Shiau, Clinical and laboratory approaches to evaluate diarrheal disorders, Crit. Rev. Clin. Lab. Sci. 25 (1) (1987) 43–69.
B. Shinkins, B.D. Nicholson, T.J. James, J.N. Primrose, D. Mant, Carcinoembryonic antigen monitoring to detect recurrence of colorectal cancer: how should we interpret the test results? Clin. Chem. 60 (12) (2014) 1572–1574.
M. Simrén, P.O. Stotzer, Use and abuse of hydrogen breath tests, Gut. 55 (3) (2006) 297–303.
C.G. Sørensen, W.K. Karlsson, H.C. Pommergaard, J. Burcharth, J. Rosenberg, The diagnostic accuracy of carcinoembryonic antigen to detect colorectal cancer recurrence - a systematic review, Int. J. Surg. 25 (2016) 134–144.
J.R. Strosberg, Diagnosis of the carcinoid syndrome and tumor localization, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA, 2019 (accessed 11.01.18).
L. van Dam, E.J. Kuipers, M.E. van Leerdam, Performance improvements of stool-based screening tests, Best. Pract. Res. Clin. Gastroenterol. 24 (4) (2010) 479–492.
M. Vives-Pi, S. Takasawa, I. Pujol-Autonell, R. Planas, E. Cabre, I. Ojanguren, et al., Biomarkers for diagnosis and monitoring of celiac disease, J. Clin. Gastroenterol. 47 (4) (2013) 308–313.
N.V. Voskoboev, S.J. Cambern, M.M. Hanley, C.D. Giesen, J.J. Schilling, P.J. Jannetto, et al., Fecal electrolyte testing for evaluation of unexplained diarrhea: validation of body fluid test accuracy in the absence of a reference method, Clin. Biochem. 48 (16–17) (2015) 1126–1130.
M.J. Yeung, J.L. Pasieka, Gastrinomas: a historical perspective, J. Surg. Oncol. 100 (5) (2009) 425–433.

GI CHEMISTRY—Self-assessment questions

1. The most cost-efficient, reliable test for *H. pylori* is:
   - **a.** Culture
   - **b.** *H. pylori* PCR
   - **c.** *H. pylori* stool antigen
   - **d.** Serum antibody to *H. pylori*
   - **e.** Urea breath test
2. In a patient with signs and symptoms of malabsorption, the best initial test to document malabsorption is:
   - **a.** Carotene in serum with the patient on usual diet
   - **b.** D-Xylose test
   - **c.** Fecal fat excretion
   - **d.** Lactose tolerance test
   - **e.** Schilling test
3. Which of the following tests is considered the best test for diagnosis of celiac disease?
   - **a.** Antibody to gliadin
   - **b.** Antibody to tissue transglutaminase
   - **c.** D-Xylose test
   - **d.** Fecal fat excretion
   - **e.** Lactose tolerance test
4. Which of the following is considered a biomarker for inflammatory bowel disease (IBD)?
   - **a.** Gliadin
   - **b.** Neutrophil cytoplasmic antigens
   - **c.** Nuclear antigens
   - **d.** Calprotectin
   - **e.** C-reactive protein
5. Which of the following can be a contributing factor to *C. difficile*-related diarrhea?
   - **a.** Prolonged antibiotic use
   - **b.** Chronic use of proton pump inhibitor drugs (PPIs)
   - **c.** Use of probiotic supplements
   - **d.** A and B
   - **e.** B and C
6. Currently, screening for colorectal cancer is usually done by means of:
   - **a.** Colonoscopy yearly
   - **b.** Fecal DNA tests yearly
   - **c.** Fecal occult blood tests yearly
   - **d.** Fecal occult blood tests yearly, followed by colonoscopy or sigmoidoscopy if abnormal
   - **e.** X-ray examination of the abdomen

Answers

1. c
2. c
3. b
4. d
5. d
6. d

Chapter 33

Evaluation of exocrine pancreatic function

Steven W. Cotten

Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Learning objectives

After reviewing this chapter, the reader should be able to:

1. State the physiological actions of the pancreas.
2. Differentiate endocrine and exocrine pancreatic function.
3. List the enzymes secreted by the pancreas and their functions.
4. State the methodology used in assays for enzyme determination in the clinical lab.
5. Describe the expected biochemical profile for relevant pancreatic disorders.
6. Recognize the laboratory's role in validating off-label use of FDA-cleared assays for pancreatic function evaluation.

Anatomy and physiology of the pancreas

The pancreas is an organ 12−15 cm in length located along the posterior wall of the abdominal cavity. The organ's shape consists of a "head," nestled in the duodenal curve that extends from the stomach, and a "tail" extending left toward the spleen (Fig. 33.1). The head of the pancreas is connected to the duodenal loop via the hepatopancreatic ampulla (also known as the ampulla of Vater). The union of the bile duct and pancreatic duct creates the hepatopancreatic ampulla, and the release of bile and pancreatic juice into the duodenal loop is controlled by the Sphincter of Oddi. The pancreas is one of just two organs having both endocrine and exocrine functions, with the other being the liver. Endocrine glands secrete hormones directly into the circulation, while exocrine glands secrete compounds to epithelial cells via a series of ducts.

The endocrine functions of the pancreas are largely related to glucose homeostasis and carbohydrate metabolism. The cells forming the Islets of Langerhans in the pancreas are responsible for the synthesis and secretion of insulin, glucagon, pancreatic polypeptide, and somatostatin. (Table 33.1) The endocrine functions of the pancreas and carbohydrate metabolism are covered in Chapter 34: Carbohydrate disorders. In certain rare cancers, endocrine cells of the pancreas can also produce other peptides, such as gastrin (Zollinger−Ellison syndrome) and vasointestinal peptide (Verner Morrison syndrome).

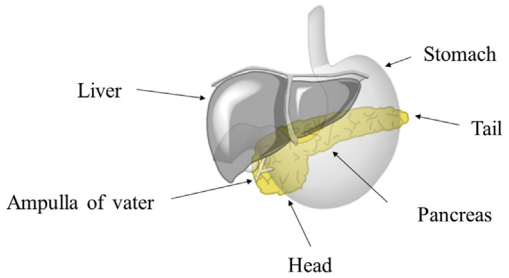


FIGURE 33.1 Anatomy of the pancreas.

The exocrine functions of the pancreas relate to the secretion of enzymes or factors that control the degradation of proteins, lipids, and polysaccharides in the duodenum and gastrointestinal tract. These enzymes include amylase, lipase, trypsin, chymotrypsin, carboxypeptidase, elastase, and phospholipase A2 (Table 33.2). These enzymes, in concert with bile salts, allow for the breakdown, emulsification, and reabsorption of sugars, amino acids, and lipids. Impaired enzyme production or secretion by the pancreas results in maldigestion and malabsorption of these breakdown products.

Laboratory measurement of pancreatic enzymes

Amylase

α-Amylase (EC 3.2.1.1), also known as 1,4-α-D-glucan glucanohydrolase, catalyzes the hydrolysis of 1,4-α-glucosidic linkages between adjacent glucose residues within complex carbohydrates. Amylase is present in several organs and tissues with the highest concentration in the salivary glands. In addition to salivary glands and the pancreas, amylase activity has been detected in semen, testes, ovaries, fallopian tubes, striated muscle, lungs, and

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00033-8

TABLE 33.1 Endocrine pancreatic secretions.

Hormone	Expression	Biological role
Insulin	β-cells	Reduces blood glucose
Glucagon	α-cells	Increases blood glucose
Pancreatic polypeptide	F-cells	Regulates endocrine and exocrine pancreatic secretions and acts as an antagonist to cholecystokinin
Somatostatin	δ-cells	Inhibits insulin and glucagon release
Gastrin	Zollinger–Ellison syndrome	Increases gastric acid, pepsinogen, secretin, and intrinsic factor
Vasointestinal peptide	Verner Morrison syndrome	Stimulates bicarbonate secretion and inhibits gastrin stimulated gastric acid secretion

TABLE 33.2 Exocrine pancreatic secretions.

Enzyme	Substrate
Amylase	1,4-α-glucosidic linkages in polysaccharides
Lipase	Glycerol esters of long-chain fatty acids
Trypsin (1 and 2)	Peptide bonds formed by the carboxyl groups of Lys or Arg with other amino acids
Chymotrypsin (1 and 2)	Peptide bonds involving carboxyl groups of Trp, Leu, Tyr, or Phe, with preference for the aromatic residues
Carboxypeptidase (A and B)	C-terminal peptide and proteins
Elastase-1	Carboxyendopeptidase that catalyzes hydrolysis of native elastin, with a special affinity for the carboxyl group of Ala, Val, and Leu
Phospholipase A-2	Phospholipids

adipose tissue. The salivary glands produce amylase (S-type) encoded by the *AMY1* gene that initiates the breakdown of starch in the mouth and esophagus. Salivary amylase content varies, but, on average, may account for up to 66% of circulating plasma amylase [1]. Genomic analysis has revealed a correlation between low copy number of *AMY1* genes and a predisposition to obesity [2,3].

The pancreas produces amylase (P-type) encoded by the *AMY2A* and *AMY2B* genes, and the enzyme is secreted into the duodenum via the pancreatic duct system. The primary physiological role of pancreatic amylase is to facilitate the breakdown of polysaccharides present in the chyme in the duodenum. Pancreatic amylase does not show a significant difference in reference intervals between males and females during development. Both sexes show a slight increase during development, with upper ranges for pancreatic amylase activity starting at approximately 28 U/L and moving toward 44 U/L by age 17 [4].

Macroamylase refers to a circulating complex of amylase and immunoglobulins (usually IgA, followed by IgG) [5]. The combined molecular weight of amylase ($\sim$62 kDa) and immunoglobulins (360-kDa IgA dimer) prevents clearance by the kidneys, resulting in accumulation of macroamylase molecules in the plasma. Most cases of macroamylasemia are idiopathic and have little clinical significance. Amylase measurements usually reveal elevated amylase activity in the absence of acute pancreatitis. In the cases of suspected macroamylasemia, polyethylene glycol (PEG) precipitation or gel filtration chromatography may be performed [6]. For PEG precipitation, the sample is treated with a PEG solution that precipitates large antibody complexes. The solution is then centrifuged to pellet any insoluble proteins and the supernatant is measured for remnant amylase activity. If macroamylase molecules were present, the residual activity of amylase should be decreased. The cutoff used for

TABLE 33.3 Amylase and lipase substrates used in common automated chemistry platforms.

Instrument platform	Amylase substrate	Lipase substrate
Beckman coulter AU	CNP-G_3	1,2-Diglyceride
Beckman coulter synchron	Ethylidene-4-NP-G_7	DGGMR
Ortho diagnostics	Amylopectin	1-Oleoyl-2,3-diacetylglycerol
Roche Cobas/Integra	Ethylidene-4-NP-G_7	DGGMR
Siemens Advia	Ethylidene-4-NP-G_7	DGGMR
Siemens Vista	CNP-G_3	DGGMR

4-NP-G_7, 4,6-Ethylidene-4-nitrophenyl-α-(1→4)-D-maltoheptaoside; *CNP-G_3*, 2-chloro-4-nitrophenyl-α-D-maltotrioside; *DGGMR*, 1,2-o-dilauryl-rac-glycero-3-glutaric acid-(6′-methylresorufin) ester.

determination is dependent upon the specific procedure used [5–8].

Determination of total amylase activity

Most methods for amylase measurement involve the assessment of enzyme activity rather than total amylase protein concentrations. Conventional assessment by the clinical laboratory measures both salivary and pancreatic amylase activities in the specimen. The majority of modern methods use defined substrates that generate colorimetric products upon digestion by amylase enzymes (Table 33.3). The unique reactivity of amylase toward each individual substrate imparts different specificities for each assay. This can impact reference intervals and interassay comparisons. The International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) published recommendations for standardization of amylase activity measurements at 30°C in 1999 and at 37°C in 2006 [9,10]. An earlier published recommendation for amylase measurement used the substrate 2-chloro-4-nitrophenyl-α-D-maltotrioside (CNP-G_3) (Fig. 33.2).

In the original IFCC method, amylase catalyzes the cleavage of the CNP-G_3 substrate liberating 2-chloro-4 nitrophenol (CNP), resulting in an absorbance change that is monitored at 410 nm. The release of CNP can be either from the original CNP-G_3 substrate or from the intermediate product CNP-G_2. No auxiliary enzymes are required for the reaction. The most recent IFCC recommendation uses the substrate 4,6-ethylidene-4-nitrophenyl-α-(1→4)-$_D$-maltoheptaoside (ethylidene-4NP-G_7). In this reaction, the substrate is protected with the ethylidene group on the nonreducing end of the oligosaccharide. The moiety improves the stability of the substrate by preventing slow hydrolysis from α-glucosidase, an auxiliary reagent. The reaction works in a similar fashion to the CNP-G_3 substrate by measuring the liberation of 4-nitrophenol that is measured via a color change at 410 nm. Using assays traceable to the most current IFCC method, serum reference intervals for adults are reported to be 31–107 U/L. Notable differences between the substrates used in the aforementioned methods are readily observed on proficiency testing participant summaries between peer groups. Samples near the upper limit of the reference interval of 107 U/L show variation between 87 and 134 U/L across all peer groups.

Commutable reference material for α-amylase

External quality assessment and multilaboratory studies have demonstrated wide interlaboratory variation of the catalytic measurement of α-amylase in serum samples. Currently, there are two primary reference measurement procedures (PRMPs) for amylase that occupy the highest position in the traceability chain. These methods use certified reference material (CRM) IRMM/IFCC-456 that consists of purified pancreatic amylase protein in an artificial matrix. As the supply of IRMM/IFCC-456 material diminishes, a new CRM is sought that may also offer commutability for calibration of routine methods. Recently five candidate reference materials were evaluated for commutability for α-amylase using clinical and laboratory standards institute (CLSI) EP30-A and the bias method recommended by the IFCC Working Group on Commutability [11]. Four of the five candidate reference materials contained amylase purified from human pancreatic tissue, and the other candidate reference material contained recombinant amylase. Three of the four candidate reference materials containing purified amylase materials were stabilized in buffer and one of four was stabilized in human serum. Two candidate reference materials containing purified amylase from human tissue showed acceptable commutability with Biosystems, Roche, and Siemens amylase assays. Differences in commutable acceptability were observed depending upon the statistical method (CLSI or IFCC) chosen; the IFCC

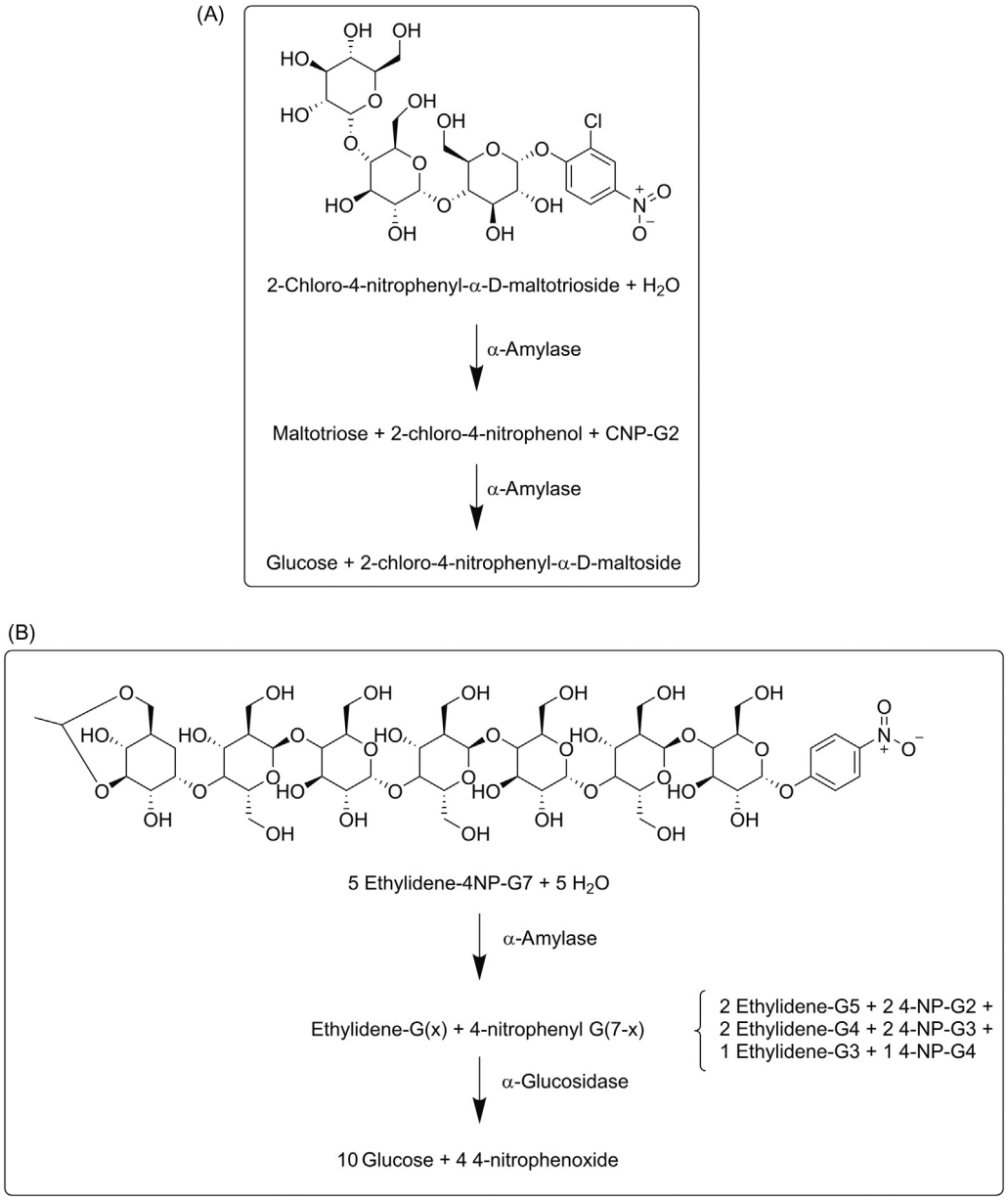


FIGURE 33.2 Two distinct substrates used for amylase activity measurement. (A) Chloronitrophenol coupled to three glucose molecules. (B) Ethylidene 4-nitrophenol coupled to seven glucose molecules.

method employs a more stringent benchmark for agreement. The evaluation of a candidate reference material shows that commutability is achievable for routine methods in an artificial matrix, but further work is required to generate a universal reference material across all amylase methods and substrates.

Determination of pancreatic-specific amylase activity

Total amylase measurement has limited specificity due to the contributions of both pancreatic and salivary amylase. The exclusive measurement of pancreatic amylase provides improved diagnostic accuracy for the identification of patients with suspected acute pancreatitis. Specific measurement of pancreatic amylase activity involves the use of two monoclonal antibodies that bind and inhibit the salivary amylase enzyme in vitro. After incubation of patient sera with these immunoinhibitory antibodies, the uninhibited pancreatic amylase can be measured using the substrates described earlier. Using pancreatic-specific amylase reagent, the IFCC reference interval has been reported for adults as 13−53 U/L [10]. Prior to immunoinhibition for specific measurement of pancreatic amylase

activity, amylase isoenzymes were separated using electrophoresis [12,13]. The type of matrix used for separation (cellulose acetate, agarose, or polyacrylamide) substantially impacted the degree of protein resolution. In addition to the separation of pancreatic and salivary amylase, other posttranslational modifications, such as glycosylation or deamination, could be observed in the form of three bands in both the P- and S-type amylase migration windows.

Lipase

Lipases (EC 3.1.1.3) refer to a class of enzymes that catalyze the hydrolysis of the glycerol esters present in long-chain fatty acids (triglycerides) into β-monoglyceride and two free fatty acids (Fig. 33.3). Several types of lipases have been identified, such as lipoprotein lipase, hepatic lipase, and pancreatic lipase. Pancreatic lipase shows selectivity for long-chain fatty acids, while other ancillary lipases, such as lipoprotein lipase and hepatic lipase, are more specific for short-chain fatty acids. The concentration of pancreatic lipase is approximately 9000 times greater in the pancreas as compared to other tissues. Complete enzyme activity is achieved only when the enzyme is present in an emulsion, usually facilitated by bile salts, such as sodium deoxycholate, and in the presence of the cofactor colipase. The bile salts promote and stabilize the emulsion of the long-chain fatty acid substrates for pancreatic lipase. Colipase binds to pancreatic lipase resulting in a conformational change of the complex that increases the binding affinity and catalytic activity of the enzyme. In addition, the presence of colipase and bile salts in the reaction mixture mitigates lipoprotein lipase activity in vitro, which increases the analytical specificity of lipase assays. All major diagnostic manufacturers have lipase reagent formulations supplemented with colipase and at least one bile acid, which may be listed as a surfactant or stabilizer in the product insert.

Determination of lipase activity

Currently there is no established consensus for a pancreatic lipase reference method [14]. Modern methods for lipase measurement utilize a variety of synthetic substrates whose breakdown products are measured. The most common substrate utilized by automated chemistry instruments is 1,2-o-dilauryl-rac-glycero-3-glutaric acid-(6'-methylresorufin) ester (DGGMR; see Table 33.3) Under basic conditions, pancreatic lipase, in concert with colipase and bile salts, cleaves the conjugated glutamic acid and phenoxazine dye from the triglyceride (Fig. 33.4). Base-mediated separation of glutaric acid from the phenoxazine liberates the dye and results in a color change that is measured at 560 nm. The measured rate of color change is proportional to the amount of lipase present in the sample. Using this substrate, reference intervals for lipase in a normal adult population are approximately 15−51 U/L, but may vary slightly based on specific population and study design.

Several diglyceride substrates have also been employed for lipase measurement due to their superior solubility [15,16]. However it has been suggested that these methods may be more prone to interferences from nonpancreatic lipase enzymes. Recently a new substrate, 1,2-dioleoylglycerol (DODG), has been reported and proposed as another candidate reference method for lipase [15]. Several modifications have been interrogated to improve the selectivity toward pancreatic lipase, including changing bile salts from sodium deoxycholate to a combination of sodium taurodeoxycholate and sodium glycodeoxycholate, a reduction in the measurement pH from 8.15 to 7.0, and an increase in temperature from 37°C to 43°C. The IFCC and Japanese Society of Clinical Chemistry (JSCC) are both actively working toward defining a reference method and reference material; however, pancreatic lipase remains one of the only enzymes that is not yet standardized in the clinical laboratory [17].

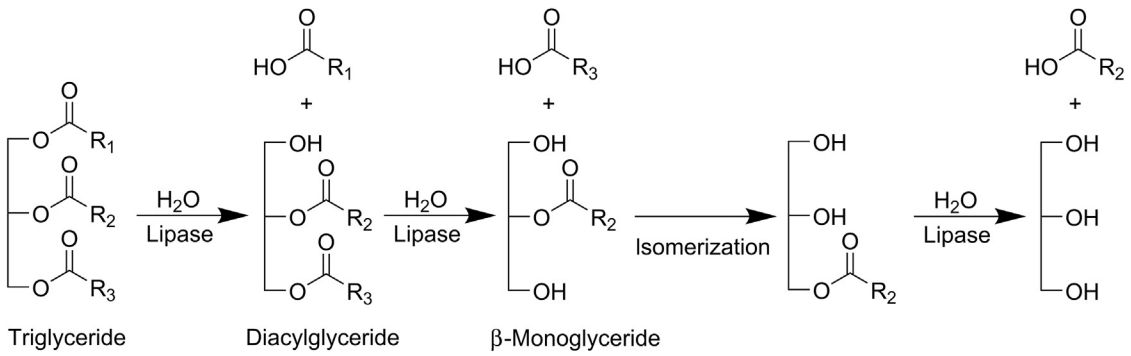


FIGURE 33.3 The breakdown of endogenous glycerol esters present in triglycerides by pancreatic lipase.

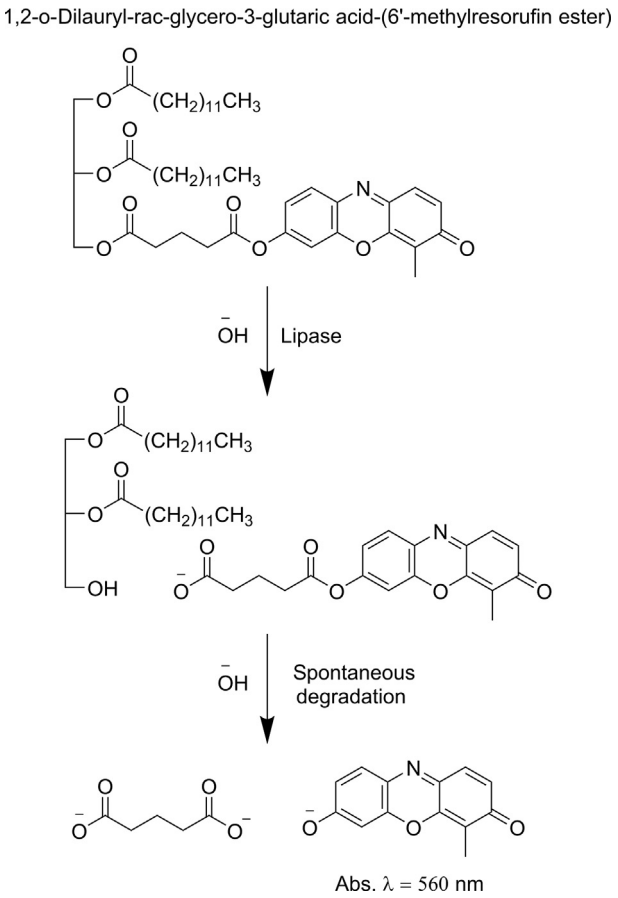


FIGURE 33.4 The enzymatic reaction scheme for the measurement of lipase activity using the substrate 1,2-o-dilauryl-rac-glycero-3-glutaric acid-(6'-methylresorufin) ester.

Trypsin

Trypsin (EC 3.4.21.4) is a serine protease enzyme produced by the pancreas that cleaves carboxyl groups present on lysine and arginine. Three isoforms (trypsin 1, 2, and 3) are synthesized and stored in the pancreas as inactive trypsinogens. Upon stimulation by cholecystokinin, the pancreas releases inactive trypsinogens into the intestine where they are activated by enterokinase via proteolytic cleavage. Protease inhibitors, such as α-1-antitrypsin (A1AT) and β-2-microglobulin (B2M), bind activated trypsin that may enter the circulation as a means to protect plasma proteins. The measurement of trypsin complexed to an inhibitor (immunoreactive trypsin) has limited clinical utility for patients with acute pancreatitis [18]. Immunoreactive trypsinogen (IRT) measurement is used extensively in newborn screening programs for cystic fibrosis, but is not diagnostic for the disease [19].

Chymotrypsin

Chymotrypsin (EC 3.4.21.1) is another serine protease produced by the pancreas that hydrolyzes the peptide bonds of tryptophan, leucine, tyrosine, and phenylalanine. Two forms of chymotrypsin (1 and 2) are synthesized by the pancreas as inactive chymotrypsinogens. Once released into the intestinal tract, the precursor enzymes are converted to chymotrypsin through proteolytic cleavage by trypsin. Chymotrypsin is more resistant to degradation than trypsin and therefore can be measured in the feces. Measurement of chymotrypsin in stool has been one method to assess adequate exocrine pancreatic function and may be useful in detecting compliance with exogenous pancreatic enzyme supplementation. Issues related to collection and analysis of fecal chymotrypsin have led to the development of alternative approaches to assess adequate pancreatic function. Collection and analysis of secretin-stimulated duodenal fluid provide a direct assessment of pancreatic function and are discussed later in this chapter.

Elastase-1

Elastase (EC 3.4.21.36) is a serine protease produced by the pancreas that catalyzes cleavage of carboxyl groups present on small hydrophobic amino acids, such as glycine, alanine, and valine. Its primary role is the breakdown of elastin, a protein that imparts elasticity to connective tissue. There are up to eight genes that encode elastase or elastase-like enzymes, four of which are classified as chymotrypsin-like. Neutrophil elastase breaks down the outer membrane of *E. coli* and other Gram-negative bacteria. Measurement of fecal elastase-1 can be used to indirectly assess exocrine pancreatic sufficiency [20]. Currently, there are two reagents commercially available for measurement of elastase with varying specificities (Table 33.4) [21]. One reagent is a monoclonal antibody against pancreatic elastase-1 and the second assay contains two polyclonal antisera mixtures against elastase. Both assays measure total elastase content, not enzyme activity. Based on their low specificities, neither reagent is recommended as a broad screening test [22]. Comparison of the two elastase reagents reveals that they detect different epitopes and the polyclonal elastase reagent reacts with nonhuman elastase. This lack of specificity of the polyclonal reagent for human elastase becomes problematic in patients with chronic pancreatitis (CP) that are receiving porcine enzyme supplements [21]. Low false positive stool elastase results can occur with small bowel bacterial overgrowth and watery stool.

Quantitative fecal fat

The measurement of fecal fat is used as another indirect measurement of pancreatic insufficiency. Conventional measurement requires the patient to be on a standard diet containing 8−12 g of fat daily for five days prior to collection of the last 72 hour stool sample. The sample is

TABLE 33.4 Reported sensitivity and specificity of elastase antisera.

	Luth et al. ($n = 127$)		Keim et al. ($n = 45$)	
	Sensitivity	Specificity	Sensitivity	Specificity
Monoclonal reagent	100	55	68.9	77
Polyclonal reagent	N.P.	N.P.	77	76

N.P., Not performed.

homogenized and fatty acids are extracted; extracts are commonly measured gravimetrically. However, this original method showed poor interlaboratory agreement and was both challenging for the patient and the laboratory [23]. Three subsequent methods have been reported that seek to better harmonize results and simplify analysis. The first method uses midinfrared spectroscopy following a liquid extraction with petroleum ether or chloroform. The assay is reported to have acceptable interassay and intraassay precision with %CVs <10 [24]. The second and third methods use ^{13}C-NMR or ^{1}H-NMR spectroscopy to measure the natural abundance of either ^{13}C or ^{1}H in fatty acids [24,25]. Several improvements to these NMR methods have been published that obviate the need for homogenization, lyophilization, or extraction of the sample, thus making NMR methods more suitable for high-throughput analysis with minimal sample processing. These methods dramatically reduce sample preparation and analysis time from hours to minutes [25].

The clinical utility of fecal fat analysis from timed stool collections has been questioned extensively with many compelling arguments against its use [26,27]. Proper collection requires the patient to adhere to a moderate to high-fat diet for several days, which may be difficult for patients experiencing diarrhea. Poor analytical performance has been documented by the absence of internal quality controls and external quality assessment between laboratories. Clinical specificity toward true fat malabsorption is low and the test results do not identify the cause of possible malabsorption. For these reasons, many laboratories restrict the use of fecal fat analysis in favor of alternative assessment of malabsorption [27].

CA 19-9

CA 19-9, also known as the Sialyl-Lewis^{A} antigen, is a tetrasaccharide carbohydrate originally described in 1979 [28]. The monoclonal antibody (1116 NC 19-9) was characterized from a human colorectal carcinoma cell line and the antigen was subsequently defined as a siaylated lacto-*N*-fucopentase II oligosaccharide. CA 19-9 is synthesized by normal epithelial tissue in the pancreas, biliary ducts, colon, endometrium, and salivary glands. A small amount of CA 19-9 can be found in plasma present on large circulating mucin glycoprotein complexes. Plasma levels of CA 19-9 rise with the progression of neoplastic diseases, such as pancreatic and colorectal cancer. The automation of tumor marker analysis has greatly decreased imprecision for CA 19-9 laboratory assays, but comparison of CA 19-9 results across multiple platforms reveals large discordant values. Method comparison using Passing−Bablok analysis generates slopes ranging from 1.0−2.0 and correlation coefficients from 0.85−0.98 for five commercially available reagents [29]. The observed lack of harmonization reinforces the need to use a single assay for serial monitoring of patients.

Disorders of the pancreas

Exocrine pancreatic insufficiency

Exocrine pancreatic insufficiency refers to the inability of the pancreas to produce the digestive enzymes required for the proper breakdown of fats, sugars, and proteins. This condition can arise from most pancreatic diseases, including acute pancreatitis, chronic pancreatitis (CP), pancreatic cancer, and cystic fibrosis. It has been estimated that malabsorption only occurs after 90% of pancreatic enzyme output has been lost [30]. Clinical presentation includes abdominal bloating and cramping, loose malodorous stools, fat-soluble vitamin deficiencies, and weight loss. Laboratory evaluation of exocrine pancreatic insufficiency can be indirect through measurement of excreted enzymes and fecal fat analysis or direct measurement of collected duodenal fluid after secretin stimulation of the pancreas. Direct measurements are suitable for detecting mild to severe insufficiency, while indirect methods are only considered suitable for the detection of moderate to severe cases.

Despite its limitations, 72-hour fecal fat analysis is currently considered the "gold standard" for indirect laboratory evaluation of exocrine pancreatic insufficiency; however, fecal elastase-1 and fecal chymotrypsin testing have also been used [31]. A comparison of quantitative fecal fat analysis and elastase-1 demonstrated that

elastase-1 (at a cutoff of 200 μg/g) predicts fat malabsorption only in patients who have not undergone partial surgical resection of the pancreas [32].

Several studies have published discordant clinical sensitivities of fecal elastase-1 to detect pancreatic insufficiency based on the specific reagent used (monoclonal or polyclonal antibody-based method). One study ($n = 127$) using the monoclonal antibody reported sensitivity and specificity for severe insufficiency reported as 100% and 55%, respectively, at a cutoff of 200-μg/g stool [22]. A subsequent study ($n = 45$) comparing the monoclonal reagent with a polyclonal reagent calculated sensitivities and specificities of 68.9% and 77%, and 76% and 77%, for the monoclonal and polyclonal reagents, respectively [21].

Direct assessment of pancreatic insufficiency

Direct pancreatic function tests involve hormonal stimulation of the pancreas through IV administration of secretin, cholecystokinin, or both [31]. The resulting duodenal secretions from the pancreas are collected and analyzed for bicarbonate, pH, and lipase. Contemporary procedures collect specimens through the accessory channel of an endoscope while the patient is sedated. Bicarbonate values >80 mmol/L observed 60 min after stimulation suggest normal duct-cell function [33,34]. Lipase measurements evaluate acinar-cell function, and pH measurements are used to monitor possible contamination from gastric fluid.

Collection of duodenal fluid may be recommended as a standard practice for direct assessment of pancreatic insufficiency by gastroenterologists; however, the analysis of this fluid creates challenges for the laboratory. Duodenal fluid is typically not FDA-approved for bicarbonate, pH, or lipase measurements, and their analyses represent off-label reagent use. This effectively classifies these assays as laboratory developed tests (LDTs) and clinical laboratories should perform the appropriate validations prior to analysis. Several publications have evaluated the performance of automated chemistry analyzers for total CO_2 or bicarbonate analysis from duodenal fluid with the back titration reference method [35–37].

Acute pancreatitis

Acute pancreatitis refers to the transient inflammation of the pancreas. Pancreatitis occurs when digestive enzymes of the exocrine pancreas spill into the parenchyma and upon activation by trypsin, autodigest pancreatic tissue. Clinical presentation is limited to nonspecific symptoms, such as abdominal pain, fever, nausea, and tachycardia. The most common causes of acute pancreatitis are gallstones (40%–70%) followed by alcohol abuse (25%–35%) [38]. Pancreatitis from gallstones usually resolves when the stone is removed or is passed spontaneously. Other minor causes of acute pancreatitis may include medications [6-mercaptopurine, azathioprine, and DDI (2',3'-dideoxyinosine)], hypertriglyceridemia (>1000 mg/dL), trauma, autoimmune diseases, genetic factors, and pancreatobiliary tumors.

Diagnosis of acute pancreatitis

Guidelines from the American College of Gastroenterology for the diagnosis of acute pancreatitis require two of the three following features: abdominal pain, elevated serum lipase, or amylase >3x the upper reference interval, and characteristic findings of pancreatitis on computed tomography (CT) scan [39]. Elevation of amylase and lipase peaks within the first 24 hours. Amylase can remain elevated for 72–96 hours, while lipase can remain elevated for 7–14 days. The kinetic differences between amylase and lipase can be attributed to different serum half-lives. Lipase has a reported half-life of 6.9–13.7 hours, which is longer than that of amylase [40–42]. The prolonged half-life is attributed to filtration of lipase by the kidneys and almost completes resorption by the proximal tubules. The extent of either enzyme elevation does not directly correlate with disease severity. Lipase is the preferred enzyme due to its extended elevation but an elevation in either enzyme fulfills criteria for diagnosis. Amylase has decreased sensitivity and specificity due to its possible normalization after 24 hours and nonspecific elevation in pregnancy, renal failure, and esophageal perforation.

Assessment of disease severity

Up to eight scoring systems have been proposed to stratify severity of acute pancreatitis in the clinical setting [39]. Each scoring system evaluates a slightly different clinical question, such as organ failure, diagnosis of severe acute pancreatitis, mortality, admission to an ICU, likelihood of readmission, or diagnosis of mild acute pancreatitis.

Chronic pancreatitis

CP refers to the progressive chronic inflammation, scarring, and irreversible fibrotic damage to the pancreas that result in loss of endocrine and exocrine pancreatic function [43]. Clinical presentation includes abdominal pain, steatorrhea, pancreatic calcifications, and diabetes. This can be caused by recurring episodes of acute pancreatitis or slow subclinical damage that occurs over many years (as seen in cystic fibrosis). The etiology of CP can be classified using the TIGAR-O system: toxic metabolite (alcohol, tobacco, hypercalcemia, hyperlipidemia, chronic

renal failure, medications, and toxins), idiopathic (early onset, late onset, and tropical), genetic (*PRSS1, CFTR*, and *SPINK1*), autoimmune (isolated, syndromic), recurrent and severe (post necrotic, vascular/ischemic, post irradiation), and obstructive [pancreas divisum, sphincter of Oddi disorders, duct obstruction (tumor), and posttraumatic pancreatic duct scars] [44].

The incidence of CP is estimated to be 4−12 per 100,000 individuals per year [44]. The incidence is higher in men than women by a factor of 1.5−3. Smoking and alcohol are both independent risk factors, and, concomitantly, they are multiplicative in the risk for the development of CP. Alcohol is the single most common risk factor and accounts for 44%−65% of CP cases. A threshold of four to five alcoholic drinks per day clearly increases the likelihood of developing CP. The prevalence of pancreatitis for alcoholics is three to six times greater than that for nondrinkers. Tobacco use (<1 pack per day) increases the risk of CP 2.4-fold; this risk increases to 3.3-fold when tobacco use increases to >1 pack per day. Smoking cessation decreases the risk of developing CP by 50% and reduces the risk of pancreatic calcifications in active disease.

Diagnosis of chronic pancreatitis

Recently, the American Pancreatic Association (APA) proposed a diagnostic algorithm for the diagnosis of CP (Fig. 33.5) [44]. The algorithm integrates both direct and indirect pancreatic function tests as well as imaging modalities, such as CT, magnetic resonance imaging (MRI), and endoscopic ultrasound. Clinical signs of CP, including history, physical exams, and existing laboratory data, initiate the workup. Fecal elastase measurement is suggested if not already performed. Fecal fat may be performed, but due to its cumbersome collection, challenging laboratory analysis, and patient dietary restrictions, its utility is diminished. Both elastase and fecal fat are moderately sensitive for the detection of pancreatic insufficiency in established CP, but may not detect or differentiate mild or early disease. These indirect tests are followed up by CT scans evaluating calcifications or atrophy. If inconclusive, then MRI imaging followed by endoscopy ultrasound is recommended. Inconclusive MRI or endoscopy may be followed by direct analysis of pancreatic function (secretin stimulation pancreatic function tests). If bicarbonate values do not meet criteria for diagnosis (<80 mmol/L at 60 min) or are inconclusive, the last remaining diagnostic test is endoscopic retrograde cholangiopancreatography (ERCP). This imaging modality, which requires contrast dye administration, should be reserved only for patients with whom direct pancreatic functions tests are equivocal.

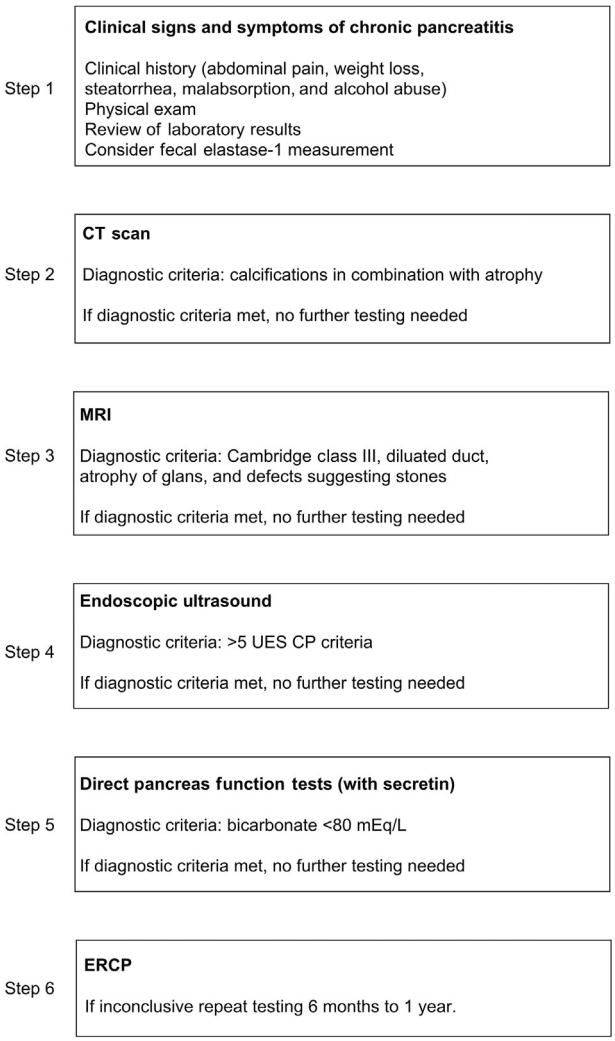


FIGURE 33.5 Diagnostic algorithm for the diagnosis of CP.

Pancreatic cancer

Pancreatic cancer is one of the most lethal cancers due to the typical late-stage diagnosis of the disease. Approximately 90% of those diagnosed with pancreatic cancer die from the disease within 5 years [45]. Smoking, long-standing diabetes, heredity factors, CP, and obesity are large contributing risk factors for the development of pancreatic cancer. Approximately 60%−70% of pancreatic tumors are located at the head of the pancreas. Early detection and surgical resection can increase 5-year survival to 20%−30% [46]. Currently, there is no effective screening tool for asymptomatic patients or early stage disease.

Imaging modalities, such as CT, MRI, and endoscopy ultrasound, are the primary diagnostic resources for pancreatic cancer. CA 19-9 is the only marker that may be used in symptomatic patients. A review of the performance of CA 19-9 has a reported median sensitivity of

79% (70%–90%) and median specificity of 82% (68%–91%) [47]. There are several limitations to the use of CA 19-9 as a biomarker for pancreatic cancer that contribute to its low sensitivity and specificity. Of note, 5%–10% of Caucasians cannot synthesize the CA 19-9 epitope due to a deficiency in fucosyltransferase [48]. Thus negative results will be obtained in these patients, even in the presence of advanced disease. Transient CA 19-9 elevations may occur in nonmalignant disorders, such as inflammation, obstructive jaundice, CP or acute pancreatitis, liver cirrhosis, and cholangitis. The frequency of obstructive jaundice in hepatobiliary disease (50%), acute pancreatitis (75%), and CP (20%) diminishes the specificity of CA 19-9. Additionally, other unrelated conditions, such as heart failure, Hashimoto's thyroiditis, rheumatoid arthritis, and diverticulitis, have all been reported to elevate CA 19-9.

If CA 19-9 is found to be elevated in the setting of pancreatic cancer confirmed by imaging studies, it may be useful in staging/prognosis and assessment of chemotherapy response [48]. The degree of elevation of CA 19-9 can be used to stratify patient survival. Both preoperative and postoperative measurements have been suggested to be predictive of patient survival and resectability of the tumor. Several studies have also demonstrated that serial measurements of CA 19-9 during the course of chemotherapy correlate with disease progression. Results suggest that changes in measurements over time correspond to tumor growth as well as overall survival [49–53].

Pancreatic cyst fluid analysis

Pancreatic cysts are detected in approximately 1%–3% of patients who undergo abdominal imaging for an unrelated condition [54]. These cysts can be classified into inflammatory cysts (pseudocysts), cystic neoplasms, and true cysts. The correct classification of the type of cyst greatly impacts treatment with the clinical goal being the identification and resection of cysts that may progress to invasive carcinomas. In addition to imaging, fine needle aspiration (FNA) followed by biochemical analysis of the cyst fluid and cytology can help differentiate the type of cyst. The analysis of cyst fluid for glycoproteins, such as CA 19-9, CA 125, CA 15-3, and CA 72-4, seeks to classify mucinous from nonmucinous neoplasms. Cyst fluid CA 19-9 values >50,000 U/mL have a reported sensitivity of 86% and specificity of 85% in distinguishing cystadenocarcinomas from other cystic lesions [55]. Cyst fluid CEA measurement has been shown to be the most accurate method to differentiate mucinous from nonmucinous lesions. Accuracy of cyst fluid CEA is 79% compared with endoscopy ultrasound (51%) and cytology (59%) using a cutoff of 192 ng/mL [56].

In addition to the measurement of tumor markers, biochemical analysis of enzymes is sometimes requested on cyst fluid. Several studies have come to different conclusions about the sensitivity of these enzymes in the classification of cyst pathology or differentiation of malignancy [56,57]. Amylase measurement is suggested to have the highest specificity for differentiation of malignant cysts from nonmalignant or indeterminate lesions based on cytology. Amylase measurements were found to be lower in patients with malignancy as compared to nonmalignant lesions, with mean measurements of 55.9 U/L (malignant) versus 78.8 U/L (indeterminate) and 101.1 U/L (nonmalignant). Lipase measurements were found to be similar across all three groups, limiting its clinical utility. Laboratories should be mindful that the analysis of tumor markers in cyst fluid are laboratory developed tests and require additional validation. The viscous nature of cyst fluid and limited volumes available can be challenging on automated chemistry analyzers. Conservation of cyst fluid is imperative if both biochemical and cytological analyses are requested.

Conclusion

The endocrine and exocrine functions of the pancreas are key regulators in carbohydrate and nutritional metabolism whose dysfunction can contribute to a range of pathologies. Laboratory analysis of exocrine pancreatic function comprises a wide variety of analytes and specimen types. Many of these less common specimen types, such as stool, duodenal fluid, and pancreatic cyst fluid, have additional preanalytical and analytical considerations for proper handling and interpretation. Furthermore, analysis of routine enzymatic analytes and tumor makers in serum may be complicated due to the lack of assay harmonization and poor clinical specificity. For these reasons, it is crucial for the laboratory to play a significant role in the stewardship, education, and research for diagnostic testing related to exocrine pancreatic function.

References

[1] W. Junge, B. Troge, G. Klein, W. Poppe, M. Gerber, Evaluation of a new assay for pancreatic amylase: performance characteristics and estimation of reference intervals, Clin. Biochem. 22 (1989) 109–114.

[2] M. Falchi, et al., Low copy number of the salivary amylase gene predisposes to obesity, Nat. Genet. 46 (2014) 492–497.

[3] D. Carpenter, et al., Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes, Hum. Mol. Genet. 24 (2015) 3472–3480.

[4] S.M. Clifford, A.M. Bunker, J.R. Jacobsen, W.L. Roberts, Age and gender specific pediatric reference intervals for aldolase, amylase, ceruloplasmin, creatine kinase, pancreatic amylase, prealbumin,

and uric acid, Clin. Chim. Acta Int. J. Clin. Chem. 412 (2011) 788–790.
[5] T. Moriyama, et al., Laboratory and clinical features of abnormal macroenzymes found in human sera, Biochim. Biophys. Acta 1854 (2015) 658–667.
[6] S.P. Wyness, J.J.H. Hunsaker, S.L. La'ulu, L.V. Rao, W.L. Roberts, Detection of macro-creatine kinase and macroamylase by polyethylene glycol precipitation and ultrafiltration methods, Clin. Chim. Acta Int. J. Clin. Chem. 412 (2011) 2052–2057.
[7] R.W. Forsman, Macroamylase: prevalence, distribution of age, sex, amylase activity, and electrophoretic mobility, Clin. Biochem. 19 (1986) 250–253.
[8] G.J. Lawson, Prevalence of macroamylasaemia using polyethylene glycol precipitation as a screening method, Ann. Clin. Biochem. 38 (2001) 37–45.
[9] IFCC methods for measurement of catalytic concentration of enzymes. Part 9. IFCC method for alpha-amylase [1,4-alpha-D-glucan 4-glucanohydrolase, EC 3.2.1.1]. International Federation of Clinical Chemistry. Clin. Chim. Acta Int. J. Clin. Chem. 281, S5–S39 (1999).
[10] International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), et al., IFCC primary reference procedures for the measurement of catalytic activity concentrations of enzymes at 37 °C, Clin. Chem. Lab. Med. 44 (2006) 1146–1155.
[11] L. Deprez, et al., Commutability assessment of candidate reference materials for pancreatic α-amylase, Clin. Chem. 64 (2018) 1193–1202.
[12] T. Takeuchi, T. Matsushima, T. Sugimura, T. Kozu, T. Takeuchi, A rapid, new method for quantitative analysis of human amylase isozymes, Clin. Chim. Acta Int. J. Clin. Chem. 54 (1974) 137–144.
[13] B.K. Gillard, Quantitative gel-electrophoretic determination of serum amylase isoenzyme distributions, Clin. Chem. 25 (1979) 1919–1923.
[14] M. Panteghini, The never-ending search of an acceptable compromise for pancreatic lipase standardisation, Clin. Chem. Lab. Med. 50 (2012) 419–421.
[15] Y. Iizuka, et al., 1,2-Dioleoylglycerol method for pancreatic lipase catalytic activity in serum, Clin. Chem. Lab. Med. 50 (2012) 475–481.
[16] Y. Iizuka, et al., Improvement and evaluation of a 1,2-dioleoylglycerol method for measuring pancreatic lipase catalytic activity in serum, Clin. Chem. Lab. Med. 52 (2014) e171–e173.
[17] IFCC - eNewsletter 2013 January–February. <http://www.ifcc.org/ifcc-communications-publications-division-(cpd)/ifcc-publications/enewsletter/enews-volumes/2013-enewsletter/enewsletter-2013-jan-feb/activities-of-the-ifcc-committee-on-reference-systems-of-enzymes-(c-rse)/> (accessed 7.01.16).
[18] A. Borgström, A. Lasson, K. Ohlsson, Patterns of immunoreactive trypsin in serum from patients with acute abdominal disorders, Scand. J. Clin. Lab. Invest. 49 (1989) 757–762.
[19] M.K. Sontag, K.B. Hammond, J. Zielenski, J.S. Wagener, F.J. Accurso, Two-tiered immunoreactive trypsinogen-based newborn screening for cystic fibrosis in Colorado: screening efficacy and diagnostic outcomes, J. Pediatr. 147 (2005) S83–S88.
[20] J.S. Leeds, K. Oppong, D.S. Sanders, The role of fecal elastase-1 in detecting exocrine pancreatic disease, Nat. Rev. Gastroenterol. Hepatol. 8 (2011) 405–415.
[21] A. Schneider, B. Funk, W. Caspary, J. Stein, Monoclonal versus polyclonal ELISA for assessment of fecal elastase concentration: pitfalls of a new assay, Clin. Chem. 51 (2005) 1052–1054.
[22] S. Lüth, et al., Fecal elastase-1 determination: 'gold standard' of indirect pancreatic function tests? Scand. J. Gastroenterol. 36 (2001) 1092–1099.
[23] J.H. Van De Kamer, Quantitative determination of the saturated and unsaturated higher fatty acids in fecal fat, Scand. J. Clin. Lab. Invest. 5 (1953) 30–36.
[24] B.S. Jakobs, M. Volmer, M.T. Hofs, D.W. Swinkels, Enhanced time-saving extraction procedure for the analysis of fecal fat by Fourier transform infrared spectroscopy, Clin. Chem. 46 (2000) 1019–1020.
[25] N.L. Korpi-Steiner, J.N. Ward, V. Kumar, J.P. McConnell, Comparative analysis of fecal fat quantitation via nuclear magnetic resonance spectroscopy (1H NMR) and gravimetry, Clin. Chim. Acta 400 (2009) 33–36.
[26] Faecal fat: time to give it up. Ann. Clin. Biochem. **38**, 164–167 (2001).
[27] M. Korostensky, et al., Elimination of 72-hour quantitative fecal fat testing by restriction, laboratory consultation, and evaluation of specimen weight and fat globules, J. Appl. Lab. Med. 3 (2018) 357–365.
[28] M. Herlyn, Z. Steplewski, D. Herlyn, H. Koprowski, Colorectal carcinoma-specific antigen: detection by means of monoclonal antibodies, Proc. Natl. Acad. Sci. U. S. A. 76 (1979) 1438–1442.
[29] S.L. La'ulu, W.L. Roberts, Performance characteristics of five automated CA 19-9 assays, Am. J. Clin. Pathol. 127 (2007) 436–440.
[30] E.P. DiMagno, V.L. Go, W.H. Summerskill, Relations between pancreatic enzyme ouputs and malabsorption in severe pancreatic insufficiency, N. Engl. J. Med. 288 (1973) 813–815.
[31] P.A. Hart, D.L. Conwell, Diagnosis of exocrine pancreatic insufficiency, Curr. Treat. Options Gastroenterol. 13 (2015) 347–353.
[32] L. Benini, et al., Fecal elastase-1 is useful in the detection of steatorrhea in patients with pancreatic diseases but not after pancreatic resection, Pancreatol. J. Int. Assoc. Pancreatol. IAP Al. 13 (2013) 38–42.
[33] T. Stevens, D.L. Conwell, G. Zuccaro, S.A. Lewis, T.E. Love, The efficiency of endoscopic pancreatic function testing is optimized using duodenal aspirates at 30 and 45 minutes after intravenous secretin, Am. J. Gastroenterol. 102 (2007) 297–301.
[34] T. Stevens, et al., A randomized crossover study of secretin-stimulated endoscopic and dreiling tube pancreatic function test methods in healthy subjects, Am. J. Gastroenterol. 101 (2006) 351–355.
[35] N. Zhong, et al., An automated analyzer provides clinically concordant results to manual back titration for quantitation of bicarbonate in pancreatic juice, Pancreas 40 (2011) 422–425.
[36] Z. Xiao, R. Lopez, M.A. Parsi, M. Dodig, T. Stevens, Comparison of autoanalyzer and back titration for measurement of bicarbonate concentration in endoscopically collected pancreatic fluid, Pancreas 40 (2011) 237–241.
[37] A. Pyle-Eilola, et al., Method performance of bicarbonate and electrolytes by use of duodenal fluid for assessment of exocrine pancreatic function, J. Appl. Lab. Med. 3 (2018) 56–64.
[38] S. Tenner, J. Baillie, J. DeWitt, S.S. Vege, American College of Gastroenterology, American College of Gastroenterology

guideline: management of acute pancreatitis, Am. J. Gastroenterol. 108 (1400−) (2013) 1416.
[39] D.C. Kuo, A.C. Rider, P. Estrada, D. Kim, M.T. Pillow, Acute pancreatitis: what's the score? J. Emerg. Med. 48 (2015) 762−770.
[40] N.W. Tietz, D.F. Shuey, Lipase in serum--the elusive enzyme: an overview, Clin. Chem. 39 (1993) 746−756.
[41] F. Apple, et al., Lipase and pancreatic amylase activities in tissues and in patients with hyperamylasemia, Am. J. Clin. Pathol. 96 (1991) 610−614.
[42] J.A. Lott, S.T. Patel, A.K. Sawhney, S.C. Kazmierczak, J.E. Love, Assays of serum lipase: analytical and clinical considerations, Clin. Chem. 32 (1986) 1290−1302.
[43] M.J. DiMagno, E.P. DiMagno, Chronic pancreatitis, Curr. Opin. Gastroenterol. 27 (2011) 452−459.
[44] D.L. Conwell, et al., American Pancreatic Association Practice Guidelines in Chronic Pancreatitis: evidence-based report on diagnostic guidelines, Pancreas 43 (2014) 1143−1162.
[45] D.P. Ryan, T.S. Hong, N. Bardeesy, Pancreatic adenocarcinoma, N. Engl. J. Med. 371 (2014) 1039−1049.
[46] C.C. Chrystoja, et al., Pancreatic cancer, Clin. Chem. 59 (2013) 41−46.
[47] K.S. Goonetilleke, A.K. Siriwardena, Systematic review of carbohydrate antigen (CA 19-9) as a biochemical marker in the diagnosis of pancreatic cancer, Eur. J. Surg. Oncol. J. Eur. Soc. Surg. Oncol. Br. Assoc. Surg. Oncol. 33 (2007) 266−270.
[48] S. Scarà, P. Bottoni, R. Scatena, CA 19-9: biochemical and clinical aspects, Adv. Exp. Med. Biol. 867 (2015) 247−260.
[49] N.R. Maisey, et al., CA19-9 as a prognostic factor in inoperable pancreatic cancer: the implication for clinical trials, Br. J. Cancer 93 (2005) 740−743.
[50] C.G. Willett, W.J. Daly, A.L. Warshaw, CA 19-9 is an index of response to neoadjunctive chemoradiation therapy in pancreatic cancer, Am. J. Surg. 172 (1996) 350−352.
[51] T.M. Bauer, et al., Carbohydrate antigen 19-9 is a prognostic and predictive biomarker in patients with advanced pancreatic cancer who receive gemcitabine-containing chemotherapy: a pooled analysis of 6 prospective trials, Cancer 119 (2013) 285−292.
[52] U. Halm, et al., Decrease of CA 19-9 during chemotherapy with gemcitabine predicts survival time in patients with advanced pancreatic cancer, Br. J. Cancer 82 (2000) 1013−1016.
[53] H. Takahashi, et al., Serum CA19-9 alterations during preoperative gemcitabine-based chemoradiation therapy for resectable invasive ductal carcinoma of the pancreas as an indicator for therapeutic selection and survival, Ann. Surg. 251 (2010) 461−469.
[54] I. Chebib, K. Yaeger, M. Mino-Kenudson, M.B. Pitman, The role of cytopathology and cyst fluid analysis in the preoperative diagnosis and management of pancreatic cysts >3cm, Cancer Cytopathol. 122 (2014) 804−809.
[55] M.S. Bhutani, V. Gupta, S. Guha, D.I. Gheonea, A. Saftoiu, Pancreatic cyst fluid analysis--a review, J. Gastrointest. Liver Dis. 20 (2011) 175−180.
[56] C. Boot, A review of pancreatic cyst fluid analysis in the differential diagnosis of pancreatic cyst lesions, Ann. Clin. Biochem. 51 (2014) 151−166.
[57] A. Sakhdari, et al., Routine analysis of lipase in cyst fluid is unnecessary and does not help discriminate between nonneoplastic and neoplastic cystic pancreatic lesions, Pancreas 46 (2017) e25−e27.

Self-assessment questions

1. What is the biochemical action of α-amylase?
   - **a.** Breaks down triglycerides into glycerol and fatty acids
   - **b.** Degrades proteins at the N-terminal end
   - **c.** Degrades carbohydrates at the α-1,4 linkage
   - **d.** Degrades carbohydrates at the α-1,6 linkage
2. What is the biochemical action of lipase?
   - **a.** Hydrolyzes the glycerol esters present in long-chain fatty acids into β-monoglyceride and two free fatty acids
   - **b.** Breaks down starches into triglycerides
   - **c.** Breaks down triglycerides into glucose and fructose
   - **d.** Hydrolyzes peptide bonds formed by the carboxyl groups of Lys or Arg with other amino acids
3. Which of the following does not add specificity to lipase measurements?
   - **a.** The presence of colipase
   - **b.** The presence of bile salts
   - **c.** Reaction conditions in an emulsion
   - **d.** The concentration of substrate
4. Measurement of pancreatic-specific amylase activity is achieved through ____.
   - **a.** addition of antipancreatic amylase antibodies to determine salivary activity
   - **b.** use of antisalivary amylase antibodies in a sandwich immunoassay format
   - **c.** addition of antisalivary amylase antibodies to inhibit this isoenzyme activity
   - **d.** use of a pancreatic-amylase specific substrate that cannot be hydrolyzed by salivary amylase
5. The conventional standard method for indirect determination of exocrine pancreatic insufficiency is ____.
   - **a.** quantitative fecal fat
   - **b.** elastase-1
   - **c.** chymotrypsin
   - **d.** trypsin
6. Which of the following is the most common cause of chronic pancreatitis?
   - **a.** Pancreatic cancer
   - **b.** Alcohol abuse
   - **c.** Genetic mutation
   - **d.** α_1-Antitrypsin deficiency
7. What laboratory test(s) should be used to diagnose acute pancreatitis?
   - **a.** Lipase 2x the upper limit of normal in serum
   - **b.** Amylase 2x the upper limit of normal in serum
   - **c.** Lipase or Amylase 3x the upper limit of normal in serum
   - **d.** Lipase 3x the upper limit of normal urine.
8. What is not a limitation of CA 19-9 as marker for pancreatic cancer?
   - **a.** Reactivity with other related cancer antigen markers
   - **b.** Elevations in heart failure and Hashimotos thyroiditis.
   - **c.** Transient elevations from inflammation, obstructive jaundice or liver cirrhosis
   - **d.** A deficiency in fucosyltransferase in 5%−10% of Caucasians
9. What is the primary analyte used for direct assessment of exocrine pancreatic function after secretin stimulation?
   - **a.** Phospholipase A2
   - **b.** IRT
   - **c.** Bicarbonate
   - **d.** Gastrin
10. Pancreatic cyst fluid CEA analysis ____.
    - **a.** should not be performed by laboratories
    - **b.** is FDA approved by most manufacturers of CEA reagents
    - **c.** is not FDA approved and requires validation by clinical laboratories
    - **d.** is not considered an LDT

Answers

1. c
2. a
3. d
4. c
5. a
6. b
7. c
8. a
9. c
10. c

Chapter 34

Carbohydrate disorders

William E. Winter[1], **J. Peter R. Pelletier**[2], **and Neil S. Harris**[3]

[1]Department of Pathology and Pediatrics, University of Florida, Gainesville, FL, United States, [2]Department of Pathology and Anesthesia, University of Florida, Gainesville, FL, United States, [3]Department of Pathology, University of Florida, Gainesville, FL, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Identify the salient features of hypoglycemia in the clinical history of a patient.
- Differentiate type 1 diabetes, type 2 diabetes, other specific types of diabetes, and gestational diabetes.
- Diagnose diabetes using clinical and laboratory methods.
- Recommend laboratory testing to assess long-term glycemic control.
- Understand the relationship between insulin sensitivity and the metabolic syndrome.
- Diagnose inborn errors of metabolism that cause hyper- or hypoglycemia (including galactosemia).

Introduction

Disorders of carbohydrate metabolism can be grouped into those that involve glucose (i.e., hypo- and hyperglycemia) and the far less common disorders that involve nonglucose carbohydrates (e.g., galactose and fructose). Disorders that involve carbohydrate digestion or malabsorption (e.g., lactose intolerance) will be discussed in detail in Chapter 31, Laboratory diagnosis of liver disease, of this book.

Glucose measurements

The most common laboratory approaches for glucose measurement are the hexokinase and glucose oxidase (GO) enzymatic methods. Point-of-care (POC) devices commonly utilize the GO or glucose dehydrogenase (GDH) methods.

In the hexokinase method, glucose is phosphorylated to glucose-6-phosphate by hexokinase, with ATP providing the phosphate group. Glucose-6-phosphate dehydrogenase then catalyzes the oxidization of glucose-6-phosphate to 6-phosphogluconate with the reduction of nicotinamide-adenine dinucleotide phosphate or nicotinamide-adenine dinucleotide (NAD^+) to nicotinamide adenine dinucleotide + hydrogen (NADH). Fructose is also detected by this method; however, plasma fructose concentrations are typically low and not clinically significant.

The hexokinase reaction is summarized below:

$$\text{Glucose} + \text{ATP} \xrightarrow{\text{hexokinase}} \text{glucose-6-phosphate} + \text{ADP}$$
$$\text{Glucose-6-phosphate} + \text{NAD}^+ \xrightarrow{\text{G6PD}} \text{6-phosphogluconate} + \text{NADH} + \text{H}^+$$

Various drugs, hemolysis, hyperbilirubinemia, and triglyceride concentrations ≥ 500 mg/dL can interfere with the assay. Hemolysis produces a negative bias, whereas hyperbilirubinemia and hypertriglyceridemia cause positive biases in the glucose measurement. Serum blanking may be useful in minimizing some interferences.

The GO method selectively and specifically measures glucose via the oxidization of glucose to gluconic acid and production of hydrogen peroxide (H_2O_2) as a by-product. Catalyzed by peroxidase, H_2O_2 reacts with o-dianisidine to form the chromogenic product oxidized o-dianisidine, which is measured spectrophotometrically at 540 nm. By inhibiting the peroxidase reaction, increased concentrations of ascorbic acid, bilirubin, uric acid, and glutathione can result in falsely low glucose measurements. The GO reaction is depicted below:

$$\text{Glucose} + \text{O}_2 \xrightarrow{\text{glucose oxidase}} \text{gluconic acid} + \text{H}_2\text{O}_2$$
$$o\text{-Dianisidine} + \text{H}_2\text{O}_2 \xrightarrow{\text{peroxidase}} \text{oxidized } o\text{-dianisidine} + \text{H}_2\text{O}$$

In POC testing (POCT), GO methods using ferrocene are affected by ambient oxygen tension. Thus when using a ferrocene-based secondary reaction, hypoxemia or laboratory testing at elevations >5000 feet can produce falsely elevated glucose measurements, whereas hyperoxemia produces falsely low glucose concentrations.

$$\text{Glucose} + \text{FAD}^+ \xrightarrow{\text{GO}} \text{gluconolactone} + \text{FAD}$$
$$\text{FAD} + \text{ferrocene}^+ \dashrightarrow \text{ferrocene} + \text{FAD}^+$$
$$\text{ferrocene} \dashrightarrow \text{ferrocene}^+ + e\text{-(electrochemical detection)}$$

An electrode measures oxygen consumption in the GO−oxygen electrode method. A polarographic oxygen

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00034-X

electrode assesses the rate of oxygen consumption. H_2O_2 is degraded by reacting with ethanol in a reaction catalyzed by catalase or by reacting with iodine catalyzed by molybdate (depicted below):

$$H_2O_2 + C_2H_5OH — \text{catalase} \rightarrow CH_3CHO + 2\ H_2O$$
$$H_2O_2 + H^+ + I^- — \text{molybdate} \rightarrow I_2 + 2\ H_2O$$

Whole blood cannot be used in these analyses, because living cells consume oxygen. Since GO is specific for beta-D-glucose, mutarotase is typically included as a reagent to accelerate the interconversion of alpha-D- and beta-D-glucose. In the absence of mutarotase, reaction time must be prolonged to allow for spontaneous interconversion. In the absence of mutarotase and an extended reaction time, glucose levels will be underestimated.

GDH catalyzes the formation of glucono-delta-lactone and NADH, the latter of which is measured spectrophotometrically. Like the GO method, mutarotase can be used to accelerate the reaction. This assay is highly specific for glucose and shows high correlation with the hexokinase assay. The GDH reaction is depicted below:

$$\text{Glucose} + NAD^+ — \text{glucose dehydrogenase} \rightarrow D\text{-glucono-delta-lactone} + NADH + H^+$$

Used in POCT, the glucose dehydrogenase pyrroloquinoline-quinone (GDH-PQQ)-based glucose monitoring systems are not specific for glucose and can detect maltose, xylose, or galactose [1,2]. Because of this significant limitation, this methodology is not recommended in POCT devices especially in the hospital setting. Maltose, a glucose disaccharide, is employed as an osmolality regulator or stabilizing agent in various biological preparations. Xylose and galactose are used in some diagnostic tests and are present in some dietary supplements, herbs, and foods. Such positive interferences in the GDH-PQQ assay can lead to an overestimation in glucose concentrations [3]. Subsequent administration of insulin based on GHD-PQQ measured glucose levels has resulted in hypoglycemia and even death. Icodextrin, which is present in some peritoneal dialysis solutions, is converted into maltose and can, therefore, also produce a positive interference in the measurement of glucose.

$$\text{Glucose} + PQQ — GDH \text{ -- } > \text{gluconic acid} + PQQH_2$$
$$PQQH2 + \text{ferricyanide ----- } > PQQ + \text{ferrocyanide}$$
$$\text{ferrocyanide ----- } > \text{ferricyanide} + e\text{-(electrochemical detection)}$$

Common specimen sources for glucose measurement

Differences in reference intervals exist among capillary blood, serum or plasma, and whole blood glucose levels. Because the water content of whole blood is lower than serum or plasma, whole blood glucose concentrations are lower than serum or plasma values. Serum or plasma glucose concentrations are ~15% higher than those of whole blood. Blood gas analyzers take this into account by normalizing to a correction factor, but individuals with very high or very low hematocrits can have aberrant results. Because serum or plasma glucose concentrations are independent of hematocrit, these fluids are preferable for analysis over whole blood. Because of the improved stability of sodium fluoride (NaF)-treated plasma as compared with serum, plasma is often the preferred specimen type when prepared from a tube containing NaF [4]. A 2015 report recommended a tube containing NaF, potassium oxalate, citrate, and EDTA for maintaining glucose stability [5].

Capillary blood is comparable with arterial blood with regard to glucose concentrations. Arterial blood glucose is 2–3 mg/dL higher than venous blood glucose in the fasting state. After a glucose load, glucose concentrations may be ~15% higher in capillary blood than in venous blood, presumably because tissues clear glucose after a meal. In emergency, operating room (OR), or intensive care unit (ICU) settings where patients may suffer from hypotension or hypoperfusion, capillary glucose measurements from finger sticks, or the forearm are not recommended. This is because the capillary glucose level may not immediately reflect the plasma glucose level (e.g., there can be an approximate 15-minute delay with forearm measurements) and the measured capillary glucose may be reduced compared with the circulating plasma glucose level. From a regulatory point of view, POCT devices used in critical care settings (e.g., ICUs or ORs) must be Food and Drug Administration (FDA)-approved for use in critical care settings. This being said, there is no universal definition of what a critical care setting is; it is the responsibility of the hospital to define "critical care." Nevertheless, POCT devices approved for critical glucose testing are available (CLIA-waived HemoCue Glucose 201 Systems, Brea, CA; and CLIA-waived Nova Stat Strip, Waltham, MA). In a study from France involving critically ill patients, POC glucose measurements were validated by measuring glucose in arterial samples split between the POCT device and the core laboratory [6]. This could serve as a model for validating glucose measurements in critically ill patients.

Whereas urine glucose measurements are part of the routine urinalysis, urine glucose testing does not play a major role in the diagnosis or management of diabetes. Urine ketone testing (e.g., Acetest) is important to detect inadequate insulinization and impending ketoacidosis. When diabetic patients are ill with nausea, vomiting, or diarrhea, or have a blood glucose level greater than 200–240 mg/dL, urine ketones should be checked. Trace ketones can be treated by increasing the patient's oral

fluid intake. Urine ketones that are 1 + or greater may require additional acute insulin administrations (e.g., ~10% of the total daily dose of insulin). It is best that individually packaged sodium nitroprusside urine test strips be used, as test strips exposed to air degrade and lose their reactivity to ketones. There are reports where diabetic ketoacidosis (DKA) was missed, because the sodium nitroprusside strip had degraded causing a false-negative result [7]. Ketonemia with an unequivocally elevated plasma beta-hydroxybutyrate (BHB) concentration supports the diagnosis of DKA as opposed to hyperglycemic hyperosmolar state (HHS).

In addition, continuous subcutaneous glucose measurement systems are available and their use is increasing. In these systems, a GO-coated electrode is typically placed in the subcutaneous space for 3 or more days; glucose measurements may be obtained as frequently as every 10 seconds, reporting an average glucose (AG) concentration every 5 minutes. The meter reports these results in real time and stores the results for graphical interpretation or wireless transmission to a subcutaneous insulin pump (i.e., an "artificial closed-loop pancreas" system). Closed-loop systems linking an insulin-infusion pump and continuous subcutaneous glucose measurement are currently under intense study and reduce the frequency of hypoglycemia while also reducing hemoglobin A1c levels. In September 2016, the FDA approved Medtronic's MiniMed 670 G System as the first hybrid closed-loop system for patients with type 1 diabetes (T1DM).

Hypoglycemia

Clinical presentation of hypoglycemia

Clinically, hypoglycemia presents with neuroglycopenia and/or adrenergic symptoms and signs. With insufficient supply of glucose to the brain, impaired mentation, sleepiness, irritability, drowsiness, seizures, or coma may develop. Death can follow in cases of severe and/or prolonged hypoglycemia. The normal adrenergic response to hypoglycemia is the release of epinephrine from the adrenal medulla. The rise in epinephrine can lead to increased anxiety, sweating, tachycardia, and raised blood pressure. However, individuals who suffer from repeated bouts of hypoglycemia may not display such dramatic symptoms.

The diagnosis of hypoglycemia is based on three criteria: (1) identification of a low blood glucose as defined above; (2) symptoms compatible with hypoglycemia; and (3) resolution of symptoms with the administration of glucose. Biochemical hypoglycemia is generally defined as a plasma glucose of 45 mg/dL or less. Note that there is a "gray" zone of decreased blood glucose between 45 and 70 mg/dL. From a 2015 study in neonates, when plasma glucose concentrations were maintained at ≥47 mg/dL, neurodevelopmental outcomes at 2 years were normal [8]. The 2019 American Diabetes Association recommendations describe glucose concentrations between 54 and 69 mg/dL as "level 1" hypoglycemia, glucose concentrations below 54 mg/dL as "level 2" hypoglycemia, and "level 3" hypoglycemia is defined as "a severe event characterized by altered mental and/or physical status requiring assistance" [9].

Artifactual hypoglycemia must be excluded prior to proceeding to an extensive patient workup. Glucose will decline at a rate of approximately 2%−3% per hour once a sample is obtained, due to continued glycolytic activity within red blood cells. Even when collected in tubes containing the NaF additive, glucose will decline in the first hour until glycolysis is fully inhibited. The delayed effect of NaF is due to the inhibitory effect of fluoride on the penultimate step of glycolysis; fluoride inhibition of the enolase enzyme prevents the production of phosphoenolpyruvate (PEP), thereby disrupting the glycolysis cascade. Moreover, extreme leukocytosis (e.g., some cases of leukemia) can lower glucose after the sample is drawn, as the white blood cells will metabolize glucose. Activated white cells (as occurs in sepsis) will metabolize glucose at an increased rate. If whole blood glucose or capillary glucose is measured, polycythemia can produce artifactual hypoglycemia, because the plasma content of blood (e.g., "plasma crit") is reduced in polycythemic states.

Hypoglycemia can be classified as fasting or postprandial. Postprandial hypoglycemia (sometimes referred to as "reactive" hypoglycemia) can be a true occurrence in the setting of "dumping" syndromes that might follow gastric bypass. In these patients, if there is disordered gastrointestinal tract motility and a carbohydrate-rich meal rapidly delivered to the small intestine, there can be an initial, rapid rise in blood glucose possibly into the hyperglycemic range. In response to this acute hyperglycemia, there may be increased release of insulin producing transient hyperinsulinism, leading to a "reactive" hypoglycemia. This condition can be treated by limiting simple carbohydrate intake, decreasing the amount of carbohydrates eaten at any one time, and increasing the frequency of smaller meals. Five-hour oral glucose tolerance tests (OGTTs) should not be performed in search of "reactive hypoglycemia" [10−12]. Five-hour OGTTs can transiently display plasma glucose levels less than 40 mg/dL at 2 hours or later in otherwise healthy individuals. However, this "transient biochemical hypoglycemia" correlates poorly with symptoms of hypoglycemia. Many such patients are women with irregular eating habits, including skipped meals.

Fasting hypoglycemia can be a very serious medical condition that requires rapid and appropriate evaluation and treatment. Fasting hypoglycemia develops most commonly in the morning following an overnight fast, after skipped meals, and following rigorous physical activity.

Hypoglycemia can be elicited in the controlled environment of the hospital by fasting an individual for up to 72 hours. Failure to develop hypoglycemia after a 72-hour-monitored fast excludes medically important fasting hypoglycemia.

Fasting in the hospital setting for up to 72 hours may be needed to prove that fasting hypoglycemia is indeed present. During the fast, the subject is given nothing by mouth except glucose-free electrolyte solutions. Fasting is observed in inpatients in order to carefully record symptoms and signs of hypoglycemia while taking laboratory measurements of glucose-related metabolism and treating low blood glucose with intravenous (IV) glucose and/or glucagon should hypoglycemia occur. If hypoglycemia does occur during the fast, glucagon can be injected to observe the rise in glucose postinjection. An excessive rise in glucose following the injection of glucagon (e.g., the increase in glucose is $\geq 30-40$ mg/dL within ~ 45 minutes) supports the diagnosis of hyperinsulinism.

At baseline, plasma glucose is determined. Because hypoglycemic symptoms can become attenuated in some individuals affected with chronic and recurrent hypoglycemia, the blood glucose must be measured regularly (e.g., every 4 hours) at the POC and by the central laboratory. Urine samples are tested for ketones throughout the fast, as ketonuria is a normal response to fasting. Once the plasma glucose declines to below 60 mg/dL, venous blood draws can be increased to every 2 hours. The fast is continued until (1) symptoms of hypoglycemia are present; (2) the venous glucose is persistently <45 mg/dL (e.g., for at least 30 minutes); or (3) 72 hours have passed without either hypoglycemia or hypoglycemic symptoms developing. Once symptoms of hypoglycemia occur or the plasma glucose declines below 45 mg/dL, blood should be obtained for clinical studies listed below:

Test	Comment
Plasma glucose	Measured to diagnose hypoglycemia
Insulin	Measured to rule out hyperinsulinism (if elevated, measure C-peptide)
βHB and urine ketones	Assessment of ketosis
Optional measurements at the time of acute hypoglycemia	
Arterial blood gases	Measured to assess acid/base balance; acidosis is common in many disorders that cause hypoglycemia
Na^+, K^+, HCO_3^-, Cl^-, and urine pH	Measured to assess acid/base balance and anion gap; acidosis is common in many disorders that cause hypoglycemia
Drug screen/ethanol	Performed to rule out intoxication; ethanol can cause hypoglycemia
Cortisol	Measure when other causes are excluded
Growth hormone	Measure when other causes are excluded
Branched-chain amino acids	Assists in documenting the biochemical effects of hyperinsulinism; rarely needed
Lactic acid	Assessment for lactic acidosis and gluconeogenic defects
Free fatty acids	Assists in documenting the biochemical effects of hyperinsulinism or suspected fatty acid oxidation disorders
Urine dicarboxylic acids	Measured when an aminoacidopathy or fatty acid oxidation disorders are suspected

Mechanisms and causes of hypoglycemia

The product of pancreatic islet beta cells, insulin is an anabolic hormone that stimulates glucose uptake into skeletal muscle, adipose tissue, and hepatocytes. Insulin is initially synthesized as preproinsulin. Upon entry into the rough endoplasmic reticulum (RER), the presequence is cleaved and remains in the cytoplasm. Secretory granules then bud off from the RER. Within the secretory granule, proinsulin is cleaved into insulin plus C-peptide (C = connecting) by prohormone convertases Pc1/3 and Pc2, and carboxypeptidase E. For every molecule of insulin secreted by the beta cell into the portal system, one molecule of C-peptide will also be secreted. During passage through the liver, more insulin is removed than C-peptide so that the molar ratio of insulin to C-peptide in the systemic circulation is less than 1.

The effect of insulin is to lower plasma glucose concentrations. Within the liver, insulin stimulates glycolysis sparing fat for storage as triglyceride in adipose tissue. As well as the suppression of gluconeogenesis by insulin, amino acids are spared and are then available for protein synthesis. Hepatic glycogen synthesis is stimulated by insulin resulting in glucose being stored as glycogen. Once glycogen stores are filled, glucose is converted into triglyceride for further energy storage in adipose tissue.

Insulin rises in response to glucose absorption from the gastrointestinal tract. As plasma glucose concentrations rise, glucagon, the product of pancreatic islet alpha cells is suppressed. In the fasting state, insulin concentrations decline and glucagon concentrations rise. Low insulin concentrations and high glucagon concentrations stimulate gluconeogenesis, glycogenolysis, and lipolysis. Low insulin concentrations reduce glucose uptake by skeletal muscle, adipose tissue, and hepatocytes, whereas increased gluconeogenesis and glycogenolysis increase hepatic glucose output. Low insulin levels permit adipose tissue lipolysis, supplying fatty acids to the liver for ketogenesis. The resulting ketone bodies BHB and acetoacetate serve as alternative fuel sources sparing glucose.

Primary disorders of glucagon deficiency [13] and excess [14] are rare. Insulin excess is one of the most

severe and life-threatening causes of hypoglycemia. Deficient insulin action underlies most forms of diabetes.

Hypoglycemia occurs when glucose utilization exceeds glucose provision by the liver. The causes of hypoglycemia can be excess glucose demand (e.g., sepsis) or deficient hepatic supply of glucose (e.g., gluconeogenic defects or liver failure). Hyperinsulinism affects both processes: hyperinsulin causes increased glucose demand as tissue uptake of glucose is accelerated (e.g., skeletal muscle, adipose tissue, and the liver) and glucose production and release by the liver is suppressed.

Laboratory evaluation of hypoglycemia

The causes of fasting low blood glucose can be divided into hyperinsulinemic hypoglycemia and nonhyperinsulinemic hypoglycemia. The measurement of insulin at the time of hypoglycemia can help differentiate between these two causes. If the insulin-to-glucose ratio is ≥ 0.3 [(uIU/mL)/(mg/dL)], hyperinsulinemia is considered to be present. Some experts state that insulin should not be measurable in the plasma in the setting of hypoglycemia not due to hyperinsulinism.

The reference interval for fasting insulin alone should not be used to define hyperinsulinemia, because, in the cases of severe hypoglycemia, an insulin concentration within the normal range would still be excessive. Furthermore, because insulin suppresses ketosis, the absence of urine ketones in the face of hypoglycemia may also be suggestive of hyperinsulinemia. Absent urine ketones may also be reflective of defective ketogenesis, which is observed in fatty acid oxidation disorders (FAODs) in infants and children. Increased insulin concentrations will also inhibit elevations in free fatty acid (FFA) and alanine concentrations that would otherwise normally occur in a fasted state or hypoglycemic state in the absence of hyperinsulinism.

The sources of excess insulin can be endogenous or exogenous and the measurement of C-peptide can help differentiate these conditions. In the setting of hyperinsulinemic hypoglycemia, if the C-peptide is very low or undetectable, the source of insulin is likely to be exogenous, because injected insulin lacks C-peptide. If there has been repeated exposure to exogenous insulin, the subject may have developed insulin antibodies (IA).

A relative excess of exogenous insulin is a common occurrence in insulin-treated diabetic patients who skip meals, skip snacks before exercise, or exercise excessively.

In testing for exogenous insulin in insulin-naive patients (e.g., patients not previously treated with exogenous insulin), the clinical laboratorian must be aware of whether or not their insulin assay will detect the various forms of recombinant-DNA produced insulins that include intentional amino acid substitutions, an altered amino acid sequence, or the addition of myristic acid to insulin. The very rapid acting insulin analogues are lys-Pro (Humalog), aspart (Novolog) and glulisine (Apidra), while the long acting insulin analogues are glargine (Lantus) and determir (Levemer). Newer insulins continue to come to the market including insulin lispro (Admelog from Sanofi approved in December 2017) and insulin aspart (Fiasp from Novo Nordisk approved in September 2017), which includes L-arginine as a formula-stabilizing amino acid and niacinamide (vitamin B3), which speeds subcutaneous absorption.

Measuring insulin in insulin-treated patients is challenging, because most patients treated with exogenous insulin (regardless of the type of insulin, including human insulin) develop IA [15,16]. IA are found in $\sim$80% to $\sim$100% of insulin-treated patients [17–19]. IA develop after as little as 14 days of exogenous insulin treatment. IA usually increase total insulin levels, because IA act as binding proteins and both bound insulin and free insulin are measured by insulin immunoassays. Therefore total insulin measurements in IA-positive patients may not reflect free (and bioactive) insulin concentrations. However, free insulin assays are not standardized. Therefore there is no analytically robust method for measuring insulin in insulin-treated patients. Such measurements are usually misleading and may lead to diagnostic errors.

Most cases of hypoglycemia in insulin-treated patients are the result of accidental or intentional overtreatment with exogenous insulin. Unfortunately, a patient may need to be hospitalized and denied any access to insulin (other than as administered by the staff) to prove that hypoglycemia does not occur spontaneously. Consultation with psychiatry should be considered for possible Munchausen syndrome or Munchausen by-proxy syndrome. Many cases of "brittle diabetes" where glycemic control is very poor result from noncompliance with the administration of insulin. Again, hospitalization with insulin administered by staff may be required to prove that a person's diabetes can be controlled with standard doses of insulin.

When the insulin and C-peptide levels are both elevated, the source of excess insulin is endogenous. Endogenous hyperinsulinemia results from the ingestion of a medication that can stimulate pancreatic beta cells to secrete insulin (e.g., sulfonylureas or meglitinides), a beta cell tumor (i.e., insulinoma), beta-cell hyperplasia (e.g., nesidioblastosis in neonates), or an inborn error of metabolism producing neonatal hyperinsulinism. Hyperinsulinemia and hypoglycemia can be transient in neonates.

In the initial evaluation of "hyperinsulinemic-elevated C-peptide" hypoglycemia, sulfonylurea or meglitinide ingestion must be excluded by history and urine toxicology testing. Similar to the problem with detecting insulin analogues, later generation sulfonylureas may not be detected with older screening methods for sulfonylureas. Therefore the specificity of urine sulfonylurea testing

must be known to the clinician and the laboratorian to ensure that false-negative results are not reported.

If drug-induced hyperinsulinemia is excluded, the patient's evaluation will focus on the circumstances of hypoglycemia. Infants of diabetic mothers (IDMs) commonly display hypoglycemia after birth if the mother's diabetes was not well-controlled. Exposure to a hyperglycemic in utero environment elicits fetal hyperinsulinemia. When the umblilical cord is clamped following birth, glucose supply to the infant is abruptly curtailed yet hyperinsulinemia persists for some days or longer causing hypoglycemia. Sick premature infants, asphyxiated infants, and infants suffering from erythroblastosis fetalis may exhibit hypoglycemia from hyperinsulinemia. The pancreas of infants requiring partial pancreatecomy for persistent, severe hypoglycemia may demonstrate beta-cell hyperplasia that is histologically termed "nesidioblastosis." This is a histologic description and is not a diagnosis unto itself.

In an infant born of a nondiabetic mother who clinically presents with symptoms of hypoglycemia, familial hyperinsulinemia should be considered as a cause of neonatal persistent hyperinsulinemic hypoglycemia [20]. Several autosomal recessive inborn errors causing hyperinsulinemia have been reported involving: (1) beta-cell inwardly rectifier subunit of the potassium channel ($K_{ir}6.2$); (2) sulfonylurea receptor-1 (SUR1; which is part of the aforementioned $K_{ir}6.2$ potassium channel); (3) beta-cell glucokinase; and (4) glutamate dehydrogenase. These defects can be understood by examining the normal pathway of insulin secretion.

Insulin release from pancreatic beta cells involves: (1) initial glucose sensing; (2) ATP generation; (3) coupling energy production to calcium entry into the beta cell; and (4) insulin release (Fig. 34.1). Extracellular glucose enters pancreatic beta cells via the facilitative glucose transporter-2. Once inside the beta cell, glucose is phosphorylated by glucokinase, which is a specialized high-K_m hexokinase with limited expression. The enzyme is expressed in hepatocytes, beta cells, and intestinal mucosal cells. The rate of energy generation (e.g., ATP production/unit time) by the beta cell via glycolysis and the Krebs cycle is dependent on the rate of formation of glucose-6-phosphate. At higher plasma glucose concentrations (with higher interstitial glucose levels), the rate of formation of glucose-6-phosphate is greater, leading to increased insulin release (Fig. 34.2). ATP generated within the beta cell then binds to SUR1, which regulates the inwardly rectifier potassium channel $K_{ir}6.2$. Closure of this K^+ channel depolarizes the beta cell. With depolarization, calcium channels open, thereby increasing intracellular calcium that triggers insulin release as insulin-containing granules fuse with the plasma membrane. These granules also contain amylin, which is released together with insulin (and C-peptide).

If SUR1 is active in the absence of ATP, hyperinsulinemia results. Likewise, if $K_{ir}6.2$ closes in the absence of SUR1 stimulation, hyperinsulinemia ensues. Excessive glucokinase activity leads to hyperinsulinemia via an excessive rate of production of glucose-6-phosphate. Gain of function mutations in glutamate dehydrogenase

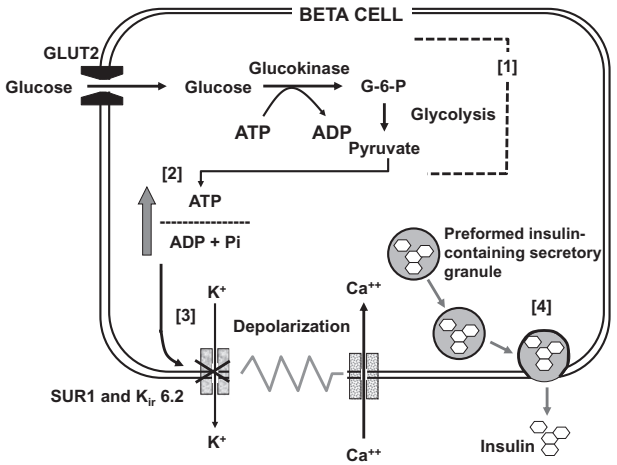


FIGURE 34.1 Insulin release is dependent on the interstitial glucose concentration. Interstitial and plasma glucose are in equilibrium and are essentially equal in concentration. Glucose enters the beta cell via glucose transporter-2. Once inside the beta cell, glucose is phosphorylated to glucose-6-phosphate as catalyzed by glucokinase, a special high Km hexokinase. Glucose-6-phosphate is converted into two pyruvate molecules that enter the mitochondrion where they undergo the Kreb citric acid cycle followed by oxidative phosphorylation and the production of ATP. An increase in the ratio of ATP to adenine dinucleotide phosphate plus Pi (inorganic phosphate) allows ATP to bind to the sulfonylurea receptor 1 closing the potassium channel ($K_{ir}6.2$). The entry of potassium into the beta cell causes beta-cell depolarization. A calcium channel opens leading to a rise in cytoplasmic calcium, which triggers insulin exocytosis.

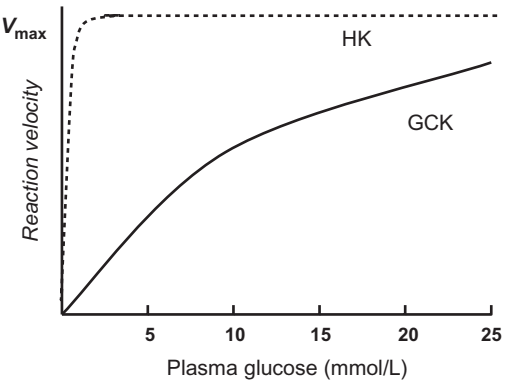


FIGURE 34.2 The high K_m glucokinase displays a rising reaction velocity in the formation of glucose-6-phosphate as interstitial glucose levels rise. A physiologic range for FPG is roughly 3.9 mmol/L (70 mg/dL) to 5.5 mmol/L (99 mg/dL). At higher glucose levels, the rate of formation of glucose-6-phosphate is increased (solid line), producing ATP at a higher rate causing greater degrees of insulin release. This is the basis for calling glucokinase the "glucose sensor" of the beta cell. Usually hexokinases display a maximum reaction velocity (V_{max}) at low glucose concentrations (*dotted line*).

produce excessive ATP production and subsequent hyperinsulinemia. Because of liver dysfunction, infants with glutamate dehydrogenase mutations can also exhibit hyperammonemia, which is a clinically useful finding in the evaluation of infants with congenital hyperinsulinism.

Insulinomas are relatively uncommon, and are diagnosed in only 1 in 1 million people per year [21]. However, the diagnosis may be difficult to establish. Some patients with proven insulinomas have had symptoms for up to 20 years. The diagnosis is based on the identification of a mass (or masses) within the pancreas using a variety of radiologic procedures, demonstration of hyperinsulinemia in a venous blood vessel draining the mass, resolution of symptoms after removal of the tumor, and immunohistological confirmation that the tumor contains insulin. Occasionally, tumors clinically suspected of secreting insulin do not stain vigorously. This may reflect recent discharge of insulin from the tumor. Although not a routine laboratory test, proinsulin levels can be disproportionally elevated in individuals with insulinomas. Proinsulin measurements are recommended by the Endocrine Society [22]; however, the basis for the recommendation is expert opinion.

Insulinomas are sometimes multifocal and not commonly malignant ($\sim$10% of all insulinomas are malignant). Malignancy in individual patients is demonstrated by the finding of metastases, for example, to the liver or regional lymph nodes. Multifocal insulinomas suggest a familial cancer syndrome such as multiple endocrine neoplasia (MEN) type 1 [23]. MEN type 1 is manifested as tumors of the pituitary, parathyroid, and pancreatic islets (the 3 "Ps"). The most common pituitary tumor is a prolactinoma. The most common pancreatic tumor is a gastrinoma. MEN type 1 is inherited as an autosomal dominant condition that results from a loss-of-function mutation in a tumor suppressor gene (menin; location: chromosome 11q13). Although localized to the nucleus, the specific function of menin remains controversial. Menin may link transcription factors and chromatin-modifying enzymes such as the histone H3K4 methyltransferase complexes (MLL1 and MLL2) [24].

If insulin concentrations are not inappropriately elevated in the setting of documented hypoglycemia, initially, a search for drugs other than sulfonylureas should be pursued. Ethanol is the most likely cause of hypoglycemia in the absence of increased insulin concentrations [25]. With the conversion of NAD^+ to NADH during ethanol oxidation, there is reduced availability of NAD^+ for the conversion of lactate to pyruvate and glutamate to alpha-keto glutarate. This results in impaired gluconeogenesis and possible hypoglycemia with fasting. Children are particularly sensitive to ethanol-induced hypoglycemia because of an increased brain-to-liver ratio (i.e., glucose consumption versus glucose production) and lower muscle mass (than adults), providing lower amounts of gluconeogenic amino acids with fasting (e.g., alanine). Other drugs causing hypoglycemia include acetaminophen via its toxic effect on the liver and propranolol, which blocks epinephrine action. Epinephrine is one of two counter regulatory hormones that protect us from acute hypoglycemia. The other hormone is glucagon. Growth hormone (GH) and cortisol provide longer-term protection from hypoglycemia by providing alternative energy sources via increased FFA production and causing insulin resistance. Cortisol also stimulates the synthesis of gluconeogenic enzymes.

Assuming that the toxicology and ethanol screens are negative, cortisol should indeed be measured. A discussion on cortisol deficiency and testing can be found in Chapter 41, Disorders of the adrenal cortex and medulla. In children, in the absence of hypocortisolism, growth hormone deficiency should next be sought (see Chapter 39: Disorders of the anterior and posterior pituitary). Deficiencies of glucagon and epinephrine essentially only occur in the setting of diabetes mellitus of 5 or more years of duration. Deficiencies of these key counter regulatory hormones may reflect some form of autonomic neuropathy. In the absence of diabetes, the rarity of epinephrine and glucagon deficiency indicates that both hormones are essential for survival.

Having excluded drugs and hormone deficiencies as causes of hypoglycemia, if not already studied, liver and renal function should be investigated using standard laboratory profiles. Any severe liver disease can cause fasting hypoglycemia, because during fasting, the liver is the single major source of circulating glucose. Renal disease is a rare cause of hypoglycemia. The kidneys have about 10% of the gluconeogenic capacity of the liver.

Hypoglycemic conditions in children

Unexplained hypoglycemia in children under age 5 should elicit a search for inborn errors of metabolism assuming other causes of hypoglycemia have been excluded.

The clinical onset of hereditary fructose intolerance occurs with the introduction of table sugar (sucrose; glucose plus fructose) into the infant's diet (Fig. 34.3). This usually occurs at nearly 6 months of age. The presentation of hereditary fructose intolerance includes vomiting postingestion of sucrose, failure to thrive, liver dysfunction,

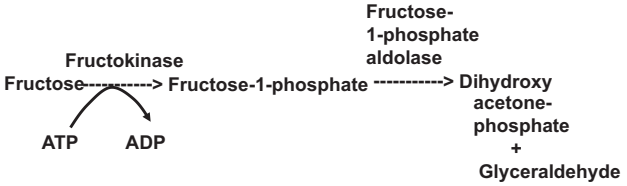


FIGURE 34.3 Benign fructosuria results from a deficiency of fructokinase. Hereditary fructose intolerance results from a deficiency of fructose-1-phosphate aldolase.

and consequent hypoglycemia. The diagnosis of hereditary fructose intolerance is made via fructose-1-phosphate-aldolase enzyme activity; chromatographic detection of fructose in urine can also support a diagnosis of fructose intolerance. Removing fructose and sucrose from the diet is the recommended treatment for hereditary fructose intolerance. Even with this dietary intervention, some children with a severe form of the disease may develop serious liver disease. Benign (essential) fructosuria results from a deficiency of fructokinase (Fig. 34.3). Fructose, sucrose, and sorbitol ingestion leads to persistent and excessive increases in blood fructose concentrations. About 10%–20% of these ingested sugars are excreted in the urine.

A cause of hypoglycemia believed to be due to substrate limitation is termed "ketotic hypoglycemia" [26]. Thin children (more commonly boys) of low birth weight between ages 18 months and 5 years who fast and develop hypoglycemia and ketosis without an identifiable inborn error are categorized as having "ketotic hypoglycemia." Insulin concentrations are appropriately suppressed in these individuals. This condition is a diagnosis of exclusion, although low alanine levels have been observed in conjunction with the hypoglycemic episode [27]. Lactate and pyruvate concentrations are within normal reference intervals. Ketotic hypoglycemia usually resolves spontaneously by 8–9 years of age.

If the liver is enlarged (with or without splenomegaly), a glycogen storage disease (GSD) should be considered [28]. Lactate and pyruvate are important measurements when considering gluconeogenic defects. GSD type 0 (zero), I (Fig. 34.4), III (Fig. 34.5), and VI (Fig. 34.6) may all present with hypoglycemia. GSD 0a is a deficiency of liver glycogen synthase encoded by the *GYS2* gene on chromosome 12p12.2 [29]. Glycogen synthase normally transfers glucose from uridine diphosphate (UDP)-glucose to glycogen's terminal branch. This disorder does not cause hepatomegaly, because excess glycogen is not produced. In the absence of increased glycogen production, fasting ketotic hypoglycemia results. Clinically, there may be muscle abnormalities, but intellectual development is usually normal. Reduced bone mineral density and short stature are also common in affected children.

In GSD type I (Von Gierke disease), glucose cannot be released from the liver, and similar to most other forms

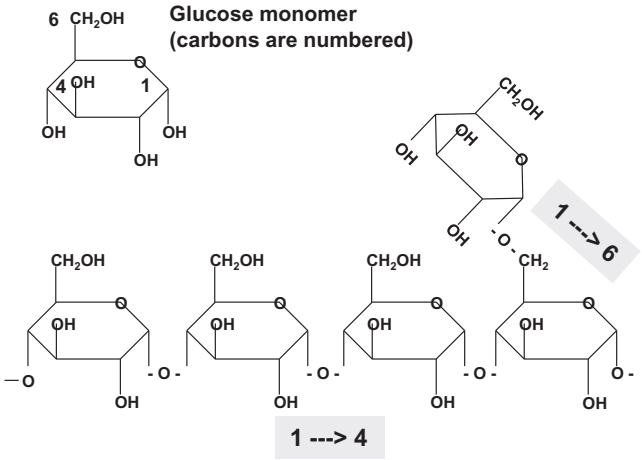


FIGURE 34.5 Glycogen storage disease type III results from a deficiency of the debrancher enzyme: amylo-[1 -> 6]-glucosidase. "1 --- > 6" depicted 1–6 branched linkages, whereas "1 --- > 4" depicts 1–4 linear linkages.

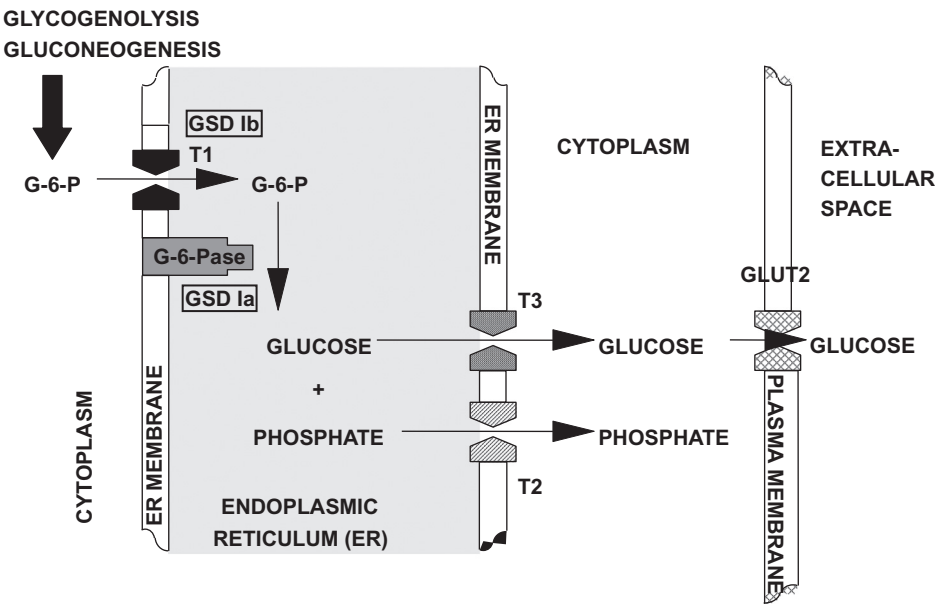


FIGURE 34.4 Glycogen storage disease types Ia and Ib are depicted. In the hepatocyte during fasting, glucose-6-phosphate results from glycogenolysis and glucoenogenesis. T1 transports glucose-6-phosphate into the endoplasmic reticulum, where glucose-6-phosphatase cleaves glucose-6-phosphate to glucose plus phosphate. T3 transports glucose into the cytoplasm, whereas T2 transports phosphate into the cytoplasm. Glucose passively leaves the hepatocyte via glucose transporter-2. *G-6-P*, Glucose-6-phosphate; *G-6-Pase*, glucose-6-phosphatase; *GLUT*, glucose transporter.

of GSD, hepatomegaly results [30]. Impaired gluconeogenesis causes hypoglycemia with fasting. There are two confirmed types of GSD type I: GSD Ia—glucose-6-phosphatase deficiency (the most common form of GSD I; the *G6PC* gene)—and GSD Ib—T1 transporter deficiency (the *SLC37A4* gene); notably, neutropenia is commonly present in the latter deficiency. With an interruption in the Cori cycle resulting from impaired gluconeogenesis in GSD type I, lactate concentrations rise, causing an anion-gap acidosis. Hypoglycemia and acidosis appear to be responsible for impaired growth in children with untreated or undertreated GSD type I. Because uric acid and lactate compete for urinary excretion via the same transporter system, hyperuricemia can result, causing gout. Since pancreatic beta cells are functioning normally, hypoglycemia suppresses release of insulin. This can lead to ketosis, elevated FFA levels, and hyperlipidemia. GSD type I is treated through nasogastric infusions of starch to prevent GSD I-related hypoglycemia [31].

GSD type I is just one type of defect that can lead to impaired gluconeogenesis. Other enzyme defects that interfere with normal gluconeogenesis include deficiencies of fructose-1,6-bisphosphatase, PEP carboxykinase, and pyruvate carboxylase (Fig. 34.7). All of these conditions can produce lactic acidosis because of impairment of the Cori cycle, thereby preventing lactate-to-glucose recycling.

GSD type III results from a deficiency of the debrancher enzyme amylo-[1 -> 6]-glucosidase. The liver and skeletal muscle are affected. GSD type VI is the consequence of phosphorylase or phosphorylase kinase deficiency. Phosphorylated phosphorylase (i.e., phosphorylase a) liberates glucose-1-phosphate residues from glycogen. Active phosphorylase kinase phosphorylates phosphorylase b to phosphorylase a.

In the setting of hypoglycemia if: (1) ketones are absent; (2) insulin is suppressed; and (3) FFA concentrations are elevated, the search for an FAOD should be undertaken. The differential diagnosis includes systemic carnitine deficiency, carnitine palmityl transferase deficiency, acyl-CoA dehydrogenase deficiency, medium chain acyl-CoA dehydrogenase deficiency, glutaric aciduria type II (multiple acyl-CoA dehydrogenase deficiencies), and hydroxymethylglutaryl coenzyme A lyase deficiency. Hypoglycemia develops in such FAODs, because ketones normally provide an

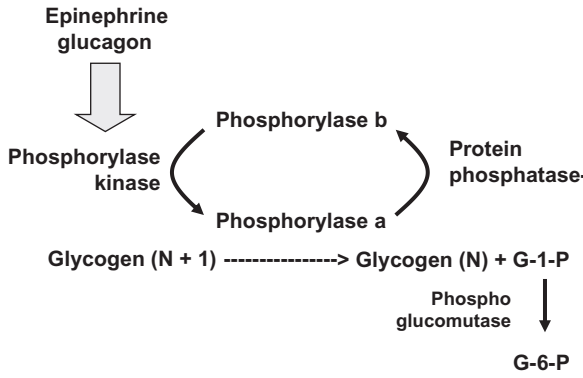


FIGURE 34.6 Glycogen storage disease type VI results from a deficiency of phosphorylase or phosphorylase kinase. The normal actions of epinephrine and glucagon are to activate phosphorylase kinase converting phosphorylase b into phosphorylase a Phosphorylase a cleaves one glucose-1-phosphate unit from glycogen. Phosphoglucomutase converts glucose-1-phosphate into glucose-6-phosphate. Protein phosphatase-1 converts phosphorylase a into phosphorylase b. *G-1-P*, Glucose-1-phosphate; *G-6-P*, glucose-6-phosphate.

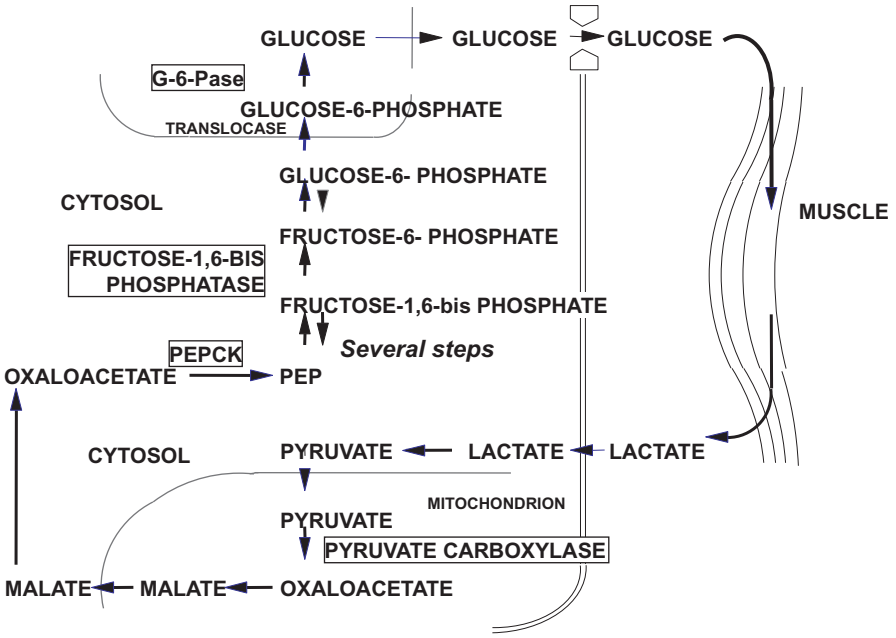


FIGURE 34.7 Gluconeogenesis is outlined. The key (irreversible) gluconeogenic enzymes are pyruvate carboxylase, phosphoenolpyruvate carboxykinase, fructose-1,6-bisphosphatase, and glucose-6-phosphatase.

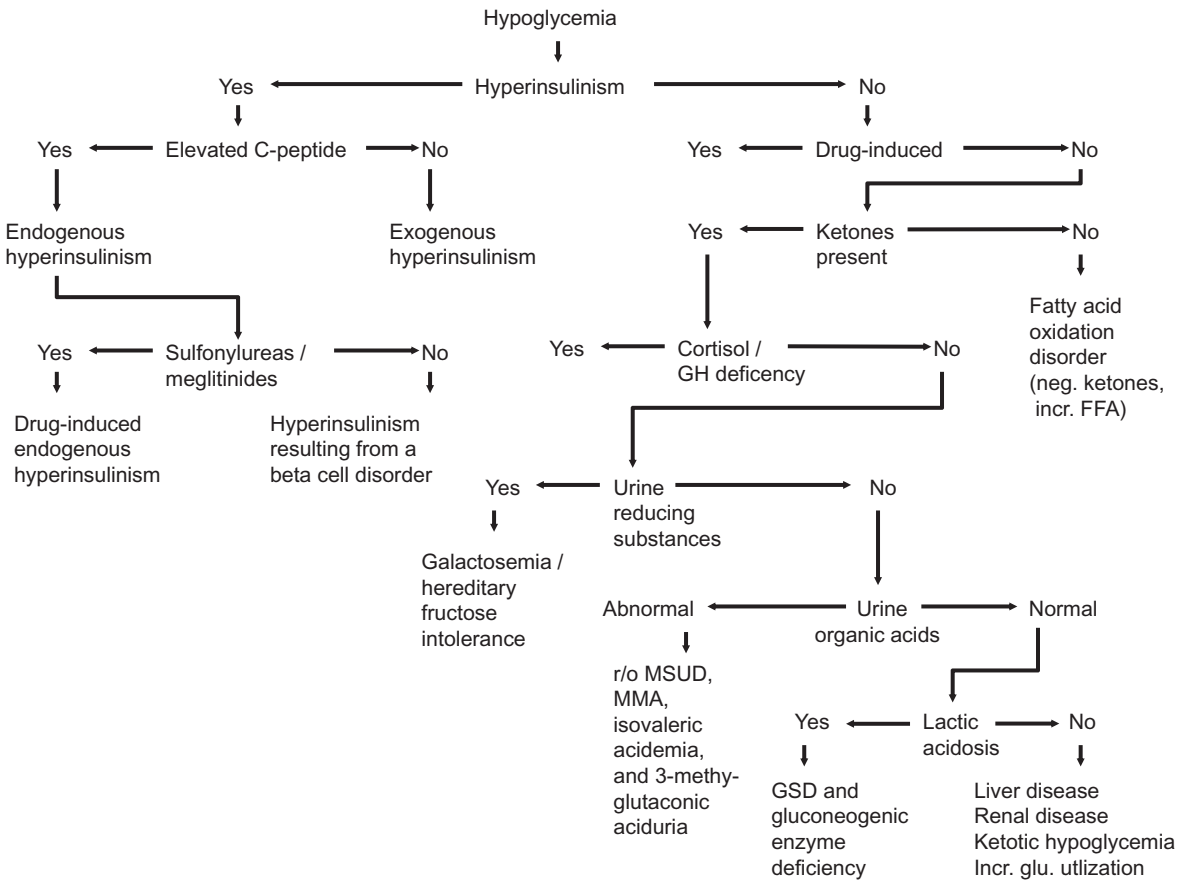


FIGURE 34.8 Overview of a reasonable approach to the evaluation of hypoglycemia. *FFA*, Free fatty acids, *GH*, growth hormone, *glu.*, glucose; *GSD*, glycogen storage disease; *incr.*, increased; *MMA*, methylmalonic aciduria; *MSUD*, maple syrup urine disease; *Neg.*, negative.

alternative energy source during times of fasting. Without these alternative energy sources, the body is more dependent on glucose for its fasting energy needs, thereby increasing the likelihood of hypoglycemia in infants and children. Screening for FAODs via mass spectrometric techniques has become routine in the United States [32]. Amino acidopathies can produce hypoglycemia, likely via impairment of hepatic function. Such amino acid disorders that can lead to hypoglycemia include methylmalonic aciduria, maple syrup urine disease, isovaleric acidemia, 3-methyl-glutaconic aciduria, and tyrosinemia type I. Specific testing for FAODs, amino acidopathies, and disorders of organic acids is beyond the scope of this chapter. Such investigations can encompass measurements of plasma amino acids, urine organic acids, ketones, lactic acid, carnitine, and FFAs.

Fig. 34.8 provides an overview of a reasonable approach to the evaluation of hypoglycemia. Testing is influenced by the age of the patient. Inborn errors of metabolism would likely not be part of the differential diagnosis of hypoglycemia in teenagers or adults. Even though liver and renal disease are listed at the terminal branch of the nonhyperinsulinemic hypoglycemia decision tree, in reality, liver function and renal function would be some of the earliest tests performed. "Increased glucose utilization" refers to conditions such as hypoglycemia occurring in children with cyanotic heart disease and hypoglycemia occurring in the setting of sepsis.

Hyperglycemia

Chronic hyperglycemia with the development of long-term complications constitutes the clinical description of diabetes mellitus. In Greek, diabetes means "siphon" referring to the polyuria that diabetic subjects suffer from. Mellitus in Greek means "sweet."

The complications of diabetes are classified as macrovascular, microvascular, neuropathic, and perinatal in cases of diabetes in pregnancy. Macrovascular disease manifests as premature and severe coronary, carotid, and/or peripheral vascular atherosclerosis. Retinopathy and nephropathy are the primary microvascular concerns associated with diabetes. Neuropathies can affect a single cranial nerve (i.e., mononeuropathies) or, more commonly, peripheral nerves (i.e., polyneuropathies). Polyneuropathies are expressed as sensory-motor deficits, painful neuropathies, or paresthesias (a "pins-and-needles" sensation). Autonomic neuropathies

can produce impotence, hypotension upon arising to the standing position, loss of beat-to-beat variation, gastroparesis (decreased intestinal motility), or "diabetic" diarrhea. Diabetes in pregnancy is discussed below in the section on "Gestational diabetes mellitus." Presently, four major categories of diabetes are recognized: T1DM, type 2 diabetes (T2DM), other specific types of diabetes, and gestational diabetes mellitus (GDM) [33].

Pathophysiologically, decreased insulin action of any cause leads to hyperglycemia. Insulin action is the product of the plasma insulin concentrations and the insulin sensitivity of target tissues. Insulin receptor variants rarely arise; thus insulin resistance is almost always a postreceptor aberration. When plasma glucose concentrations exceed 150−180 mg/dL, the renal threshold for glucose reabsorption by the renal tubules is exceeded and glycosuria develops. This is manifested as polyuria (increased urination) that results from an osmotic diuresis. Children may suffer from enuresis (bed wetting), while both adults and children can experience nocturia (urinating at night). Polydipsia (increased drinking) is a physiologic response to increased urinary fluid loss. Glycosuria can cause urinary sodium, potassium, magnesium, bicarbonate, and phosphate wasting. Hyperglycemia associated with plasma glucose concentrations ≥250 mg/dL also impairs normal neutrophil function, thereby producing an acquired defect in the innate immune system and increased risk of bacterial and fungal infections. Poor wound healing is typical in individuals with poorly controlled diabetes. This can result from impaired microvascular circulation, dehydration, acidosis, and/or neutrophil impairment.

Type 1 diabetes

T1DM results from autoimmune destruction of the pancreatic beta cells [34]. Seventy-five percent of patients develop T1DM prior to 18 years of age. While patients are usually lean at the time of onset, because the endemic frequency of obesity has reached epidemic proportions, obesity should not exclude the diagnosis of T1DM. Up to 15% of persons with T1DM will have a similarly affected first-degree relative. Northern Europeans and Caucasian North Americans are most commonly affected. The highest prevalence of T1DM is in Finland, where ~1 in 100 people are affected with T1DM. T1DM is rare in Asia. The frequency of T1DM in Japan is ~1 in 10,000 (a rate one−one hundredth of that observed in Finland).

At onset, one or more islet autoantibodies can be found in ≥95% of T1DM individuals [35]. The most common islet autoantibodies at onset are islet cell cytoplasmic autoantibodies (ICA) and glutamic acid decarboxylase autoantibodies (GADA). Each autoantibody is present in 70%−80% of new-onset patients. Insulinoma-associated-2 autoantibodies and zinc transporter 8 (ZnT8A) autoantibodies are slightly less common at onset (~60%). Insulin autoantibodies (IAA) are detected in ~50% of children at T1DM onset; however, IAA are uncommon in adults. The GADA, IA-2, IAA, and ZnT8A are measured by immunoassay, whereas ICAs are measured by indirect immunofluorescence. In the case of IAA, assays do not discriminate IAA from IA that are developed as a consequence of exposure to exogenous animal or human insulin. Therefore IAA should be measured within 14 days of the initiation of insulin therapy. After that time, IA can emerge in insulin-treated patients.

From a genetic point of view, the alleles human leukocyte antigen (HLA)-DR3, HLA-DR4, HLA-DQB1*0201, and HLA-DQB1*0302 are increased in frequency in T1DM subjects compared with the general population. These alleles are genetic variants of various class II major histocompatibility complex protein chains. Heterozygosity for HLA-DR3 and HLA-DR4 increases the risk of developing T1DM by a factor of 12. However, the absolute risk of T1DM is still "relatively" low as only 1 in 40 such heterozygotes develop T1DM. The alleles HLA-DR2 and HLA-DQB1*0602 are far less common in T1DM individuals than in controls, and therefore these alleles are considered to be protective of T1DM. Only 1 in 5000 people with T1DM carry an HLA-DR2 allele. The second most important gene locus for T1DM is the insulin gene (e.g., the variable number of tandem repeats region 5' of the insulin gene). More than 10 other genes or loci have been implicated as influencing the development of T1DM [36]. Such other loci include protein tyrosine phosphatase, nonreceptor type 22, cytotoxic T-lymphocyte-associated protein 4, and the IL-2 receptor alpha chain. At the present time, genetic testing is predominantly a research tool.

A slowly progressive variant of T1DM seen in adults that initially presents as "noninsulin dependent diabetes," where GADA or ZnT8A are positive, is termed "latent autoimmune diabetes of adulthood (LADA)" [37]. Subjects who present with a "noninsulin dependent diabetes" phenotype that carries HLA-DR3 or DR4 are likely to express an islet autoantibody (most commonly GADA or ZnT8A) and are more likely to become insulin requiring compared with DR3/DR4-negative T2DM subjects.

With poor diabetic control and consequent unrestrained gluconeogenesis in T1DM, weight loss occurs as protein is metabolized to glucose. Weight loss also results from accelerated conversion of fat into ketone bodies because of severe insulin deficiency. Polyphagia (increased eating) can be a response to weight loss. Weight loss despite increased caloric intake is characteristic of untreated T1DM. With severe insulin deficiency, ketosis is not restrained leading to an anion gap acidosis [38]. Massive hyperglycemia also results with the plasma glucose concentrations regularly exceeding 500 mg/dL. Hyperosmolarity then ensues. Early on, ketosis is evident in the finding of urinary ketones (e.g., the urine

Acetest is positive). As the plasma concentration of ketones rises (i.e., increased beta-hydroxybutyric acid and acetoacetic acid), nausea with vomiting results. Fluid loss from polyuria compounded by decreased fluid intake because of nausea and vomiting and extreme ketosis can precipitate the clinical syndrome of DKA. As fluid depletion becomes more severe, lactic acidosis from hypovolemia and impaired microvascular tissue perfusion can supervene.

Untreated DKA results in a worsening cycle of hypovolemia, acidosis, impaired perfusion, declining cardiac output, impaired mentation, coma, and death. Children can develop cerebral edema and die of foraminal herniation, where displacement of the cerebellar tonsils through the foramen magnum compresses the brain stem. DKA is treated through the administration of IV fluids and an insulin infusion that halts ketosis, corrects acidosis, and lowers plasma glucose. The etiology of cerebral edema in children is controversial [39]. However, excessive fluid administration in children should be avoided. A recent study reported: "Neither the rate of administration nor the sodium chloride content of intravenous fluids significantly influenced neurologic outcomes in children with diabetic ketoacidosis" [40].

The diagnosis of T1DM is based on the findings of severe hyperglycemia, polyuria, and polydipsia and supported by a history of polyphagia, weight loss, and ketosis. Excluding obese adolescent African Americans, where DKA can occur in the setting of untreated T2DM, DKA is pathognomonic of T1DM. At diagnosis, 30% of children with T1DM still present with DKA. If the diagnosis of T1DM is in doubt, for example, a suspected case of LADA, islet autoantibody testing should be ordered. DKA in children has a mortality of $\sim$0.5%. In adults, 10%−15% of patients with DKA will succumb because of the frequent coexistence of serious disorders of the heart, lung, liver, and/or kidney.

Measurements of BHB can be helpful in diagnosing ketoacidosis. Clinicians will usually follow BHB concentrations during the treatment of DKA. Successful treatment of DKA includes a reduction in BHB concentrations into the reference interval. Despite our enthusiasm for BHB measurements, DKA can be managed successfully without such measurements [41]. Crushed "Acetest" tables should not be used to titrate blood ketones. This nitroprusside reaction is not specific for ketones. In addition, Acetest tables are no longer manufactured.

Type 2 diabetes

Approximately 95% or more of non-GDM diabetes cases are accounted for by T1DM or T2DM. Furthermore, T2DM is 9−10 times as common as T1DM. T2DM results from a combination of insulin resistance and relative insulinopenia. T2DM is not autoimmune in etiology and affected patients are negative for islet autoantibodies (including IAA). Islet autoantibodies in persons with clinically diagnosed T2DM indicate that such individuals actually have a slowly progressive form of T1DM termed "LADA." Insulin resistance is most commonly a consequence of obesity. Until recently, T2DM was observed almost exclusively in people over age 40. However, with the epidemic of childhood and young adult obesity, T2DM can now be observed well before age 40. Women are affected more often than men predominantly because of their higher rates of obesity and lower lean body mass. A family history of T2DM is very common in affected individuals. T2DM is two or more times as common in minorities (e.g., African American, Hispanic American, Asian American, Native American, and Pacific Islander) as in Caucasians. The onset of T2DM can be very subtle. In the absence of screening programs for T2DM, up to 15% of adults with T2DM display microvascular complications at onset suggesting that dysglycemia (e.g., an abnormal glucose level) has been present for long periods of time prior to the formal clinical diagnosis of T2DM.

An unusual complication of T2DM is the development of HHS (once termed: hyperosmolar, hyperglycemic, nonketotic coma) [42]. In this condition, often under stress, severe hyperglycemia develops in a patient with T2DM. Ultimately plasma glucose levels can exceed 1000 mg/dL. There is sufficient insulin produced to avoid ketosis, yet coma develops from dehydration and hyperosmolarity. Because of severe polyuria from massive glycosuria, circulating blood volume becomes depleted. When renal blood flow declines sufficiently, urine output does begin to decline. However, this decline in urine output now only further worsens the patient's hyperglycemia as less glucose is cleared through the kidney due to glycosuria. HHS is treated by gradual rehydration and an insulin infusion to reduce the plasma glucose level. HHS has been recognized as a cause of death in obese African Americans (particularly males) at the onset of T2DM [43].

Other specific types of diabetes

Other specific types of diabetes encompass eight major subcategories: (1) genetic defects in beta-cell function (e.g., insulinopathies, hyperproinsulinemias, maturity-onset diabetes of youth, and mitochondrial diabetes) [44]; (2) genetic defects in insulin action (e.g., lipodystrophy and insulin receptor mutations observed in Rabson−Mendenhall syndrome and leprechaunism syndrome); (3) diseases of the exocrine pancreas (e.g., chronic pancreatitis, cystic fibrosis, and hemochromatosis); (4) endocrinopathies (e.g., acromegaly, Cushing syndrome, and pheochromocytoma); (5) drug- or chemical-induced diabetes (e.g., glucocorticoids, growth hormone, and pentamidine); (6) infections (e.g., congenital rubella,

meningitis, and sepsis); (7) uncommon forms of immune-mediated diabetes [e.g., type B insulin resistance (insulin receptor autoantibody-mediated)]; and (8) other genetic syndromes associated with diabetes [e.g., type A insulin resistance, myotonic dystrophy, and Down, Klinefelter, Turner, Prader-Willi, Lawrence−Moon−Biedle, Werner, Cockayne and Wolfram (diabetes insipidus−diabetes mellitus−optic atrophy−deafness) syndromes].

Genetic defects in beta cell function account for the neonatal diabetes syndromes [45]. The neonatal diabetes syndromes can result from chromosome 6q24 imprinting anomalies, K_{ATP} channel mutations involving Kir6.2 (*KCNJ11*) or SUR1 (*ABCC8*), transcription factor mutations [e.g, HNF-1alpha (MODY3); IPF-1 (MODY4); FoxP3 (the IPEX syndrome), PTF1A, GLIS3, and HNF-1beta (MODY5)], enzyme loss-of-function mutations (e.g., glucokinase and eukaryotic translation initiation factor 2-alpha kinase 3), and insulin gene mutations.

Gestational diabetes mellitus

GDM is diabetes that is diagnosed during pregnancy [46,47]. The consequences to the fetus can be significant. Maternal hyperglycemia produces fetal hyperglycemia with consequent fetal hyperinsulinemia. As described above, such IDMs are at risk for neonatal hypoglycemia when, following delivery, clamping and cutting of the umbilical cord abruptly terminate the hyperglycemic state yet hyperinsulinemia continues. IDMs are physically large, which can lead to mechanical problems with delivery (dystocia) that may necessitate caesarian section. However, IDMs are biochemically less mature than age-matched control infants. IDMs suffer higher rates of hypocalcemia, hyperbilirubinemia, and respiratory distress syndrome. Because of the rising prevalence of T2DM in women of reproductive age, diabetes screening early in pregnancy is necessary to detect undiagnosed diabetes that existed prior to pregnancy. The biochemical diagnosis of GDM will be discussed below.

If diabetes is present prior to the pregnancy and the mother's glycemic control in the first trimester is not optimal, there is a threefold increased risk of the fetus suffering a birth defect. While ~3% of all infants have a recognizable birth defect, ~10% of IDMs whose mother was not in good glycemic control exhibit a birth defect. Excellent maternal diabetic control preceding conception and continuing during the pregnancy reduces the risk of birth defects to the level seen in the general population.

Diagnosis of diabetes mellitus

In patients with classical symptoms of diabetes including DKA, HHS, or polyuria and polydipsia, the demonstration of unequivocal hyperglycemia (e.g., a random plasma glucose level of 200 mg/dL or greater) on a single occasion meets the diagnostic definition of diabetes. Other markers of hyperglycemia include: (1) a fasting plasma glucose (FPG) of 126 mg/dL or greater; (2) a 2-hour glucose result of 200 mg/dL or greater during an oral glucose tolerance test (OGTT); or (3) a hemoglobin A1c result of 6.5% or greater. However, in persons presenting with DKA, HHS, or acute symptoms, most physicians would order a random plasma glucose measurement as opposed to ordering an FPG, an OGTT, or a hemoglobin A1c.

Excluding such classical symptoms, the diagnosis of diabetes rests upon the recognition of hyperglycemia on two different occasions. Hyperglycemia is defined as follows: (1) an FPG level of 126 mg/dL or greater; (2) as part of an OGTT, a 2-hour plasma glucose level of 200 mg/dL or greater; or (3) a hemoglobin A1c level of 6.5% or greater. Any combination of hyperglycemic criteria is adequate to diagnose diabetes; however, it is preferred that the same test be repeated on different days (e.g., if the FPG was elevated initially, the second test should be an FPG measurement). In the 2020 American Diabetes Association (ADA) clinical practice recommendations, 2 glucose abnormalities on a single day were added as an option in diagnosing diabetes. Therefore the following abnormalities identified on a single day can diagnose diabetes: (1) elevated FPG and elevated hemoglobin A1c, (2) elevated 2-hour plasma glucose during the OGTT and elevated hemoglobin A1c, or (3) elevated FPG and elevated 2-hour plasma glucose during the OGTT.

Blood for plasma glucose measurement is preferably placed in a gray top tube, because within 1 hour, NaF inhibition of glycolysis is complete. If blood is placed in a red top tube for serum glucose measurement or in a green top tube for plasma glucose measurement, the glucose concentration will decline by at least 2%−3% per hour. This is problem can be minimized if the sample can be rapidly centrifuged and the plasma or serum separated (note: for serum, blood should be allowed to clot for ~10 minutes prior to centrifugation). With delayed sample delivery to the laboratory, a substantial decline in the glucose concentration can be observed potentially leading to the missed diagnosis of diabetes. If a gel separator tube is used, once the sample is spun, the plasma or serum above the gel will be stable for up to 18 hours at 24°C and stable for up to 36 hours at 4°C [48].

Oral glucose tolerance test

In nonpregnant women, men, and children, the 2-hour, 2-time point OGTT is carried out following an overnight fast of 8−14 hours. A baseline (zero time) plasma glucose level is drawn and then a 75-g oral glucose load is

administered. Time zero is when the subject begins to drink the glucose beverage. The subject should finish drinking the beverage in 5 minutes or less. However, it is recommended that the glucose beverage should not be drunk rapidly, as this can induce nausea or even vomiting (especially in pregnant women). The concentration of glucose in the beverage should not exceed 25 g/100 mL. Children are administered 1.75 g/kg to a maximum dose of 75 g of glucose. At time 2 hours, a second plasma glucose level is obtained. During the test, the subject should remain sedentary. Other than water on the morning of the test, the subject should take no food, medications, or caffeine and should not smoke. The patient should be in their usual state of health, at a stable weight, and on their usual diet. The subject should not be acutely ill and should not have been recently hospitalized as this could cause a transient abnormality of glucose tolerance and a falsely abnormal OGTT. To prepare for the OGTT, for ~3 days prior to testing, the subject should ingest a minimum of 150 g of carbohydrate per day.

Whereas insulin resistance and hyperinsulinism are the characteristics of T2DM, there is no recommendation from the ADA that insulin or C-peptide be measured in the diagnosis of any form of diabetes. Furthermore, hyperinsulinemia may not persist in long-standing T2DM as progressive beta-cell failure can occur. Only when insulin or C-peptide levels are very high (e.g., T2DM) or extremely low (e.g., T1DM) might they be of diagnostic value. In comparing patients with T1DM and T2DM, there is considerable overlap in the distributions of insulin and C-peptide levels. Because the measurement of insulin and C-peptide may lack diagnostic value, insulin or C-peptide should not be measured as a routine part of the OGTT. In addition, while there is reference interval data for fasting insulin levels, there is little data on "normal" insulin responses to oral glucose.

The normal 2-hour plasma glucose during an OGTT is less than 140 mg/dL. The OGTT has no validated utility in the management of diabetes. The cutoffs defining normal and hyperglycemia are summarized below.

Time	Normal cutoffs		Hyperglycemia cutoffs	
	mg/dL	(mmol/L)	mg/dL	(mmol/L)
0	<100	<5.6	≥126	(≥7)
2	<140	<7.8	≥200	(≥11.1)

The American Diabetes Association and various other professional societies [i.e., American Association for Clinical Chemistry (AACC)] now recommend hemoglobin (Hb) A1c measurements for screening and the diagnosis of diabetes. Hyperglycemia is present when the Hb A1c is 6.5% or greater.

Prediabetes

In the absence of hyperglycemia, an FPG level of 100 mg/dL to 125 mg/dL identifies "impaired fasting glucose (IFG)." Similarly, concerning the OGTT, a 2-hour plasma glucose of 140−199 mg/dL identifies "impaired glucose tolerance (IGT)." Hb A1c values between 5.7% and 6.4% also indicate an increased risk of developing diabetes. Collectively, these carbohydrate abnormalities are termed "prediabetes." Note however that prediabetes is not a diagnosis or a disease. Prediabetes is a descriptive term. While IFG dates from 1997, IGT is a term that has been used for over 40 years.

The significance of IFG, IGT, and Hb A1c's of 5.7%−6.4% is that affected individuals have a 25%−50% chance of progressing to T2DM unless the individual changes their lifestyle to lower their caloric intake, exercise, and lose weight. Certain medications such as metformin have also been shown to decrease the rate of progression to T2DM. However, diet and exercise are twice as successful as metformin at reducing progression to T2DM [49,50].

Diagnosis of gestational diabetes mellitus

The term GDM is now restricted to diabetes, which is first diagnosed in the second or third trimesters of pregnancy. Diabetes diagnosed in the first trimester of pregnancy most likely represents preexisting or newly developed T2DM recognizing that T1DM will rarely present during pregnancy. Pregnant women with marked obesity, a previous personal history of GDM, glycosuria, or a strong family history of diabetes should be tested for diabetes as soon as it is feasible [33]. If diabetes is not diagnosed previously or is not diagnosed in the first trimester, all women should be tested for GDM between 24 and 28 weeks using one of two possible approaches.

In the one-step approach to GDM screening, between 24 and 28 weeks gestation, a fasting, 2-hour, 75-g OGTT is carried out. GDM is diagnosed when any one of the plasma glucose cutoff levels is met or exceeded:

Time	Cutoffs	
	mg/dL	(mmol/L)
0	92	(5.1)
1	180	(10.0)
2	153	(8.5)

In the two-step approach, between 24 and 28 weeks gestation, a nonfasting, 1-hour, 50-g OGTT is carried out. If the glucose is equal to or greater than 140 mg/dL, then the fasting, 3-hour, 100-g OGTT is performed. Using a cutoff of 130 or 135 mg/dL during the 1-hour, 50-g OGTT as a trigger for the 3-hours OGTT increases sensitivity for the diagnosis of GDM.

During the fasting, 3-hour, 100-g OGTT, glucose is determined at times 0, 1, 2, and 3 hours. If the plasma glucose concentrations meet or exceed the following cutoffs at 2 or more time points, GDM is diagnosed. Either the Carpenter/Coustan cutoffs or the National Diabetes Data Group (NDDG) cutoffs can be used.

	Cutoffs			
	Carpenter/Coustan criteria		NDDG criteria	
Time	mg/dL	(mmol/L)	mg/dL	(mmol/L)
0	95	(5.3)	105	(5.8)
1	180	(10.0)	190	(10.6)
2	155	(8.6)	165	(9.2)
3	140	(7.8)	145	(8.0)

Screening for type 2 diabetes in adults and children

The ADA recommends that all adults aged 45 and above be screened for diabetes every 3 years with an FPG measurement or Hb A1c measurement [51] [note: guidelines from the American College of Obstetricians and Gynecologists (ACOG) were not available to this author as one must be an ACOG member to access their guidelines]. According to the ADA, an OGTT is an acceptable alternative test for diabetes screening, but the OGTT should not be used routinely. If there is a high index of suspicion for T2DM but the FPG or Hb A1c is not clearly elevated, the physician can elect to order an OGTT on the patient.

If an adult patient is overweight [i.e., the body-mass index (BMI) is 25 kg/M^2 or greater] and has one or more risk factors for T2DM, screening should begin prior to age 45 and/or be performed more frequently. Risk factors for T2DM include a positive family history of diabetes in a first degree relative, habitual physical inactivity, being a member of a minority group, previously identified IFG, IGT or Hb A1c of 5.7%−6.4%, history of GDM or having a baby weighing more than 9 lbs, hypertension (≥140/90 mmHg), hypertriglyceridemia (triglycerides ≥ 250 mg/dL), low HDL-cholesterol (HDL-C ≤ 35 mg/dL), or a history of vascular disease or polycystic ovarian syndrome. The American College of Clinical Endocrinology (ACCE) recommends that screening of high-risk subjects be initiated at age 30.

The ADA recommends that screening for diabetes be considered in overweight children who also have two other risk factors for diabetes such as (1) a family history of T2DM; (2) minority group ethnicity; and (3) signs of or conditions associated with insulin resistance such as hypertension, dyslipidemia, acanthosis nigricans, or polycystic ovarian syndrome. Overweight in children is defined as a BMI >85th percentile for age, weight for height >85th percentile, or weight >120% of ideal for height. The preferred screening test for diabetes in children is an FPG. Hb A1c screening for diabetes is not as sensitive in children as in adults [52]. Such children should be retested at 2-year intervals beginning at age 10 or earlier if puberty develops before age 10.

Special aspects of blood glucose testing

Self-monitoring of blood glucose (SMBG) is a vital part of diabetic care in insulin-treated patients. There is increasing data to support the use of SMBG in noninsulin-treated patients who treat their diabetes intensively [53]. Initially developed for use in clinical problem solving, the continuous glucose monitoring system is FDA-approved for routine use. Furthermore, hardware and software for closed-loop subcutaneous insulin infusions controlled according to frequent subcutaneous glucose measurements (i.e., an "artificial" pancreas) are FDA-approved for adults and children aged 7 years and above [54].

Postprandial glucose (PPG) measurements may improve outcome in pregnancy [55]. Otherwise, there are a few circumstances where PPG measurements may theoretically be helpful: (1) in patients with diabetes whose preprandial glucose levels are within the target range yet their Hb A1c levels are above target; (2) when using a drug that primarily affects the PPG level to titrate drug dose; and (3) recognition of hypoglycemia following exercise or nutritional changes. The 2019 ADA guidelines state: "... it is reasonable for postprandial testing to be recommended for individuals who have premeal glucose values within target but have A1C values above target." [and] "Postprandial glucose may be targeted if A1C goals are not met despite reaching preprandial glucose goals. Postprandial glucose measurements should be made 1−2 h after the beginning of the meal, generally peak levels in patients with diabetes" [56]. For nonpregnant adults with diabetes, the preprandial capillary plasma glucose recommendation is 80−130 mg/dL (4.4−7.2 mmol/L), whereas the peak postprandial capillary plasma glucose recommendation is <180 mg/dL* (10.0 mmol/L).

Transient hyperglycemia

An area of interest is the potential value of preventing stress-induced hyperglycemia in the setting of the ICU. Initial data from a single-institution pulmonary ICU demonstrated that improved clinical outcomes were observed in the critically ill, nondiabetic patients whose glucose was maintained within narrow limits (e.g., 80−120 mg/dL) using an IV insulin infusion [57]. Following this

report, several studies also illustrated clinical benefit resulting from "tight glycemic control" (TGC). However, the value of TGC has become controversial as additional studies showed no benefit and higher rates of hypoglycemia. A metanalysis concluded that TGC is not supported by the available data [58].

Maintenance of near normoglycemia may be beneficial to the management of the patient's fluid and electrolyte balance and to assuring optimal granulocyte function, because hyperglycemia impairs granulocyte function. The ADA has provided guidelines for inpatient management of hyperglycemia [59]. The reader is referred to this resource.

Long-term glycemic monitoring in diabetes

Glycated hemoglobin measurements

The nonenzymatic glycation of proteins is a generalized phenomenon that occurs in both diabetic and nondiabetic individuals. However, because of higher plasma glucose concentrations, glycated proteins are in higher concentration in diabetic than in nondiabetic individuals. Hemoglobin can be glycated by glucose and nonglucose sugars. Glycated hemoglobin [a.k.a, hemoglobin A1 (Hb A1)] takes three major forms: Hb A1a, Hb A1b, and Hb A1c. In Hb A1c, the N-terminal valine of the beta chain is glycated with glucose. Hb A1c represents the major glycated hemoglobin species being ~85% of Hb A1. The chemistry responsible for the formation of Hb A1c is illustrated in Fig. 34.9.

Hb A1c measurements correlate with the average plasma glucose level over the preceding 2−3 months. Glycemia over the month immediately prior to Hb A1c testing has a disproportionately larger effect on the Hb A1c measurement and is responsible for ~40% of the glycated hemoglobin concentration. The Hb A1c concentration is proportional to the extent to which blood glucose is elevated and the duration of time over which the glucose is elevated. Elevated Hb A1c predicts an increased risk of microvascular and neuropathic complications, whereas reductions in Hb A1c predict reduced risks of such complications. This is true in both T1DM, as shown in the Diabetes Control and Complications Trial (DCCT) [60], and T2DM, as demonstrated in the United Kingdom Prospective Diabetes Study [61]. Therefore the measurement of Hb A1c has become the standard of practice for the assessment of glycemic control.

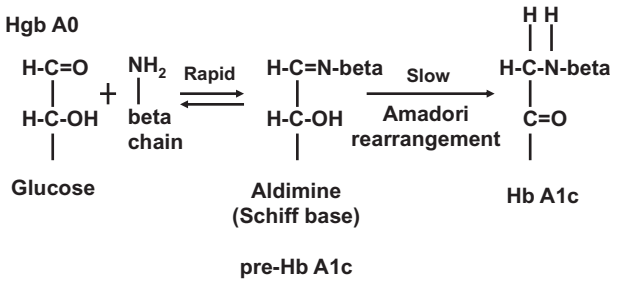


FIGURE 34.9 A rapid and reversible dehydration occurs between glucose and the N-terminus of the hemoglobin beta chain that yields an aldimine (Schiff base). This product can be termed "pre-Hb A1c." Via an Amadori rearrangement, Hb A1c is slowly and irreversibly formed.

The ADA's goal for diabetic patients is an Hb A1c of less than 7%. The 2018 guidelines state: "Providers might reasonably suggest more stringent A1C goals (such as <6.5% [48 mmol/mol]) for selected individual patients if this can be achieved without significant hypoglycemia or other adverse effects of treatment (i.e., polypharmacy)" [56]. ACCE recommends a target Hb A1c of <6.5%. Hb A1c should be measured at a minimum interval of every 6 months in patients whose Hb A1c is less than 7%. Otherwise, Hb A1c should be measured quarterly as an assessment of glycemic control.

Hb A1c can be measured using high-performance liquid chromatography (HPLC) (ion exchange or boronate affinity), gel electrophoresis, capillary electrophoresis (CE), enzymology, or, most commonly, immunoassay. At least two companies in the United States sell commercial HPLC systems designed for the measurement of Hb A1c in central laboratories [e.g., Tosoh Biosciences (South San Francisco, CA) and BioRad (Hercules, CA)]. Sebia (Lisses, France) markets a CE system for Hb A1c measurement. There are analytical and biological limitations associated with all technologies. For example, in some HPLC systems, if a patient's HbF exceeds ~15%−20%, Hb A1c cannot be accurately reported. Further, if an abnormal Hb coelutes with Hb A1c, such as Hb Raleigh, Hb Hope, Hb N-Baltimore, or Hb Franklin Park, Hb A1c cannot be accurately measured [62]. Hb A1c can be measured at the POC using manual chromatography systems or using immunoassays (e.g., A1cNow or DCA2000). The DCA2000 is a popular immunoassay POC instrument that many diabetes clinics use to measure rapidly Hb A1c in their clinic patients. Reporting Hb A1c at the time of the patient's visit has been shown to improve their clinical outcome [63].

To measure Hb A1c, a patient must have Hb A: patients lacking Hb A (e.g., Hb SS, Hb CC, or Hb SC) cannot produce Hb A1c by glycation. In addition, if erythrocyte half-life is shortened (e.g., because of hemolysis), the Hb A1c measurement may be falsely low. On the other hand, if erythrocyte half-life is elevated (e.g., because of splenectomy), Hb A1c concentrations may be increased. Recent data where Hb A1c was measured by HPLC indicated a negative Hb A1c bias of 0.3% comparing persons with sickle cell trait to controls [64]. This would result in the underdiagnosis of either prediabetes or diabetes in persons with sickle cell trait if Hb A1c were measured by HPLC. Iron deficiency raises Hb A1c on average by 1.5% producing a positive bias in Hb A1c measurements [65]. This may be because of an altered rate of glycation [66].

The original reference method for the measurement of Hb A1c was HPLC as used in the DCCT study. In the early 1990s the National Glycohemoglobin Standardization Program (NGSP) was established to develop rigorous analytic performance standards and calibrators that would standardize Hb A1c measurements throughout the United States and other areas of the world that participated in the NGSP program [67]. The NGSP used the DCCT HPLC method as its reference method for measuring Hb A1c. It was determined early on that the best way to standardize Hb A1c measurements was for manufacturers to standardize their methods using procedures and calibrator materials from the NGSP. Commercial methods that meet specific standards of precision and accuracy are declared "NGSP-certified." Individual laboratories can also apply for certification based on their analytical performance. The NGSP has progressively set more stringent precision and accuracy goals for manufacturers and laboratories that have driven improvements in assay performance.

The present reference method for Hb A1c calibration was developed by the International Federation of Clinical Chemists (IFCC) [68]. In this method, N-terminal hexapeptides are cleaved from the beta chains of hemoglobin by the action of the protease endoproteinase Glu-C. These hexapeptides are purified by reverse-phase HPLC into glycated and nonglycated hexapeptide pools. The hexapeptides are then quantified by mass spectroscopy or CE with UV detection. Calibrators are created from these materials with the Hb A1c expressed as the ratio of the glycated hexapeptide to the nonglycated hexapeptide.

For a large group of patient samples when the IFCC method was compared with the NGSP DCCT HPLC method, the NGSP method was shown to display a positive bias of $\sim$1.5%−2.0% [69]. It was subsequently decided that the IFCC method would be employed to produce calibrators used by the NGSP, manufacturers, and laboratories. However, the calibrator values would be recalculated to DCCT method results so that there was no bias from previous DCCT HPLC method results. Thus the IFCC method produced "DCCT-aligned" calibrators. The IFCC A1c method units are mmol/mole (i.e., mmol of Hb A1c per mole of Hb). In short, use of the IFCC reference method affected manufacturers and reference laboratories but did not affect the actual reporting of percent Hb A1c. In regions of the world where Hb A1c are reported based solely on the IFCC method (without reference to the HPLC HbA1c method), HbA1c is reported, as noted above, in the units of mmol/mole.

During these discussions among professional associations from North America, Europe, other parts of the world, and the NGSP, the question arose as to the relationship of Hb A1c to a patient's AG. Patients are sometimes more familiar with the concept of glucose measurements than Hb A1c measurements. This may be especially true for patients who perform SMBG. Researchers wondered how estimated average glucose (eAG) correlated with Hb A1c and if eAG should be reported together with the Hb A1c result. A 1987 publication from the DCCT [70] had shown that for every 30 mg/dL rise in the mean plasma glucose, the Hb A1c increased by $\sim$1% and that eAG and Hb A1c were linearly correlated ($R^2 = 0.64$).

In 2008 a major paper from the A1c-Derived Average Glucose Study Group [71] was published showing an even stronger correlation between Hb A1c and eAG. In their study of over 500 people with T1DM, T2DM, and normal controls, 90% of the eAG fell within $\pm$15% of the predicted estimate based on the following equation: AG mg/dL = 28.7 × A1c − 46.7. R^2 was 0.84 and the P value was <.0001. There was no effect of age, sex, diabetes type, race, ethnicity, or smoking status on this correlation. Based on this strong linear correlation, the ADA and the AACC recommend that eAG be reported along with the Hb A1c result. Below is a comparison of Hb A1c and the eAGs from the 2008 study and the 1987 DCCT study.

	2008	1987
HbA1c (%)	eAG (mg/dL)	eAG (mg/dL)
5	97	100
6	126	135
7	154	170
8	183	205
9	212	240
10	240	275
11	269	310
12	298	345

A previous controversy was the use of Hb A1c for diabetes screening and diagnosis [72]. This was resolved in 2010 as noted in the present recommendations. Advantages of Hb A1c measurement are that it can be determined regardless of the time of the day, Hb A1c is stable, and Hb A1c is relatively standardized from one laboratory to the next as a benefit of NPSG's efforts to standardize Hb A1c assays. Hb A1c is more reproducible than FPG measurements and Hb A1c is much more reproducible than 2-hour OGTT glucose values. This being said, Hb A1c testing is the least clinically sensitive of the three methods for the detection of diabetes: the 2-hour plasma glucose on OGTT testing is most sensitive (and most difficult to obtain) following by FPG measurements. In practical terms, the best way (most sensitive way) to detect hyperglycemia is the 2-hour OGTT. However, the 2-hour OGTT is the most invasive and difficult test to perform. The next best method of detecting hyperglycemia is the measurement of FPG. HbA1c measurements are the least sensitive tests for the detection of hyperglycemia; however, measuring HbA1c is the most convenient test for patients, because fasting is not required.

Glycated serum protein measurements

Proteins other than hemoglobin will become glycated in patients with diabetes. Because the major serum protein is albumin, the "glycated serum protein" test predominantly assesses the concentration of glycated albumin. The addition of glucose to albumin forms a 5-membered ring with a structure similar to fructose. This is why glycated serum albumin is usually referred to as "fructosamine." Because the half-life of albumin is ~10 days, fructosamine measurements reflect 2–3 weeks of glycemic control. Besides its use in patients that lack Hb A, fructosamine measurements may be sought during pregnancy when a shorter term integrated assessment of glycemia is required. This does not exclude also measuring Hb A1c every 2–3 months during pregnancy. During pregnancy, there may be a negative bias in Hb A1c measurements due to the development of anemia.

Future therapies for type 1 diabetes

In T1DM, there may be opportunities to apply immunotherapy to mitigate autoimmune beta cell destruction. Therefore several studies are underway developing therapies to preserve C-peptide secretion in new-onset T1DM patients, as well as prevent T1DM. These efforts have focused on suppressing or altering the immune response using putative autoantigens, drugs, or monoclonal antibodies. Efforts at impairing T cell proliferation [i.e., mycophenolate mofetil plus an anti-CD25 monoclonal antibody (daclizumab)] [73], producing tolerance with glutamic acid decarboxylase-alum [74] or insulin [75] and suppressing inflammation with an anti-IL-1 monoclonal antibody (canakinumab) or an anti-IL-1R monoclonal antibody (anakinra) [76], were not beneficial. However, B cell inhibition with an anti-CD20 monoclonal antibody (rituximab) [77] and impaired T cell activation with CTLA-4-Ig (abatacept) [78] were beneficial to preserving C-peptide secretion in new-onset T1DM individuals. Studies using anti-CD3 monoclonal antibodies, such as otelixizumab, have produced beneficial results in some studies whereas least one study showed no benefit [79–81]. Another monoclonal antibody, teplizumab (a humanized monoclonal antibody IgG1κ directed against CD3), in multiple studies showed preservation of beta-cell function compared with placebo [82–86].

Metabolic syndrome

Metabolic syndrome is a collection of disorders that are caused by, or highly associated with, insulin resistance. Insulin resistance has been strongly linked to central obesity, especially, and visceral obesity [87]. A combination of elevated FFA levels delivered to the liver from the intestine and omentum, cytokines (e.g., tumor necrosis factor-alpha in skeletal muscle), disordered adipokine secretion (e.g., a deficiency of adiponectin), and possible excess cortisol levels (from the adipose-tissue conversion of cortisol precursors to cortisol) produce insulin resistance in skeletal muscle, adipose tissue, and the liver. Fatty infiltration of skeletal muscle, liver, and the beta cells is well-recognized pathologies in this condition. Consequent to decreased insulin action, the beta cells attempt to compensate initially producing hyperinsulinemia. The combination of insulin resistance and/or hyperinsulinemia is associated with the common features of the metabolic syndrome including hypertension, dysglycemia (e.g., IFG, IGT, T2DM, and GDM), hypertriglyceridemia, low HDL-C, dense LDL, increased concentrations of plasminogen-activator inhibitor-I (which is pro-thrombotic), hyperuricemia causing gout, acanthosis nigricans, nonalcoholic fatty liver, and nonalcoholic steatohepatitis. In women, hyperandrogenism, oligomenorrhea, amenorrhea, and infertility can be observed as a consequence of the polycystic ovarian syndrome. The many components of the metabolic syndrome are highly atherogenic, such as, diabetes, hypertension, and dyslipidemia. Because if its atherogenicity, cardiologists (as well as other physicians) may refer to the metabolic syndrome as the "cardiometabolic syndrome."

Galactose disorders

Galactosemia is not a single disorder (Fig. 34.10). Disordered galactose metabolism is observed in three inborn errors of metabolism that are inherited as autosomal recessive diseases: (1) galactokinase deficiency; (2) galactose-1-phosphate uridyl transferase deficiency; and (3) uridine diphosphate galactose-4-epimerase deficiency.

The reactions that occur as part of normal galactose metabolism are summarized below:

1. Galactose + ATP <--galactokinase--> Galactose-1-phosphate + ADP
2. Galactose-1-phosphate + UDP-glucose <--galactose-1-phosphate uridyl transferase--> UDP-galactose + Glucose-1-phosphate

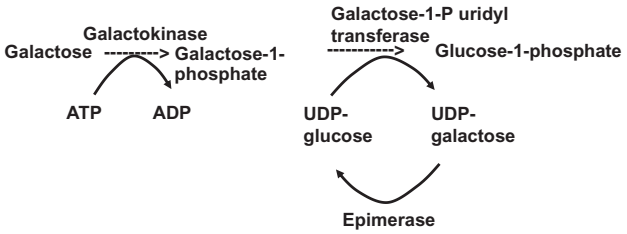


FIGURE 34.10 Galactosemia most commonly results from a deficiency of galactose-1-phosphate uridyl transferase. Galactokinase deficiency produces galactosuria and cataracts.

3. UDP-galactose <--uridine diphosphate galactose-4-epimerase--> UDP-glucose

Galactokinase deficiency, affecting 1 in ~40,000 infants, produces elevated serum galactose levels that lead to galactosuria. Aldose reductase converts galactose into galactitol. Excess galactitol draws water into the lens of the eye leading to cataract formation. Other than cataracts, galactokinase deficiency appears to be a benign condition. Galactokinase deficiency should be considered in any infant or child with cataracts.

Galactosemia, affecting 1 in ~48,000 infants, presents in the first weeks to months of life with hypoglycemia, failure to thrive, liver dysfunction (e.g., hyperbilirubinemia and hepatomegaly), and cataracts (from elevated galactose concentrations). Deficiencies of either galactose-1-phosphate uridyl transferase or uridine diphosphate galactose-4-epimerase are very serious. Transferase deficiency leads to elevated galactose with formation of cataracts. Galactose-1-phosphosate is also elevated. Galactose-1-phosphate is believed to be the toxic agent that damages the central nervous system (CNS), liver, and kidney in individuals with either the transferase or epimerase deficiency. Hypoglycemia may occur after lactose or galactose ingestion with glucose concentrations as low as 10 mg/dL. Elevated galactose-1-phosphate may interfere with glycogenolysis and gluconeogenesis. Sepsis has been reported in ~10% of cases of infants affected with galactosemia. Causes of sepsis include *Escherichia coli, Klebsiella, Enterobacter, Staphylococcus*, beta-streptococcus, and *Streptococcus faecalis*.

With untreated transferase deficiency, CNS injury can produce brain damage and serious development delay. Without a galactose-free diet, liver disease is progressive leading to fatty infiltration, pseudoacini formation, macronodular cirrhosis, and even ascites. Renal tubular dysfunction is manifested in a secondary distal (type 1) renal tubular acidosis, proteinuria, and amino aciduria. There are clinical variants of the transferase deficiency where dysfunctional alleles may have up to 50% of normal activity (such as the "Duarte" allele). Rarely, transferase activity is absent in the peripheral blood but clinical disease does not develop. This extraordinary condition may only occur in African Americans where hepatic and intestinal transferase activities are 10% of normal. Overall, transferase deficiency affects 1 in ~50,000 people in the United States.

Benign and malignant forms of epimerase deficiency have been reported. In the benign condition, epimerase is absent from the blood cells, but epimerase is present in other tissues. The serious form is similar to transferase galactosemia although cataracts have not been reported with epimerase deficiency or are far less common than in transferase deficiency. The treatment of all forms of galactosemia involves lactose and galactose restriction. Without treatment, transferase or epimerase deficiency can be fatal. For galactokinase and transferase deficiencies, such dietary restriction should be absolute and lifelong. With epimerase deficiency, a very limited amount of galactose is allowed in the diet.

Concerning neonatal screening, if the galactosemia screen is based on recognizing elevated galactose levels, several feedings must be provided before elevated galactose may be detectable. Screening for enzymatic defects is usually limited to detection of transferase deficiency. Such an assay is not dependent on feedings but will miss galactokinase and epimerase deficiencies. In summary, the diagnosis of galactosemia is based on a positive Benedict's test for reducing substances in the urine, a negative urine test for glucose by dipstick (e.g., the GO method), and an enzyme assay demonstrating reduced activity. Molecular analysis is available [88].

References

[1] K.M. Floré, J.R. Delanghe, Analytical interferences in point-of-care testing glucometers by icodextrin and its metabolites: an overview, Perit. Dial. Int. 29 (4) (2009) 377–383.

[2] V. Mathew, A. Ramakrishnan, R. Srinivasan, K. Sushma, G. Bantwal, V. Ayyar, Erroneous glucose recordings while using mutant variant of quinoprotein glucose dehydrogenase glucometer in a child with galactosemia, Indian J. Endocrinol. Metab. 17 (Suppl. 1) (2013) S289–S291.

[3] J.P. Frias, C.G. Lim, J.M. Ellison, C.M. Montandon, Review of adverse events associated with false glucose readings measured by GDH-PQQ-based glucose test strips in the presence of interfering sugars, Diabetes Care 33 (4) (2010) 728–729.

[4] M. Fobker, Stability of glucose in plasma with different anticoagulants, Clin. Chem. Lab. Med. 52 (7) (2014) 1057–1060.

[5] G. Dimeski, K.S. Yow, N.N. Brown, What is the most suitable blood collection tube for glucose estimation? Ann. Clin. Biochem. 52 (Pt 2) (2015) 270–275.

[6] J. Allardet-Servent, M. Lebsir, C. Dubroca, M. Fabrigoule, S. Jordana, T. Signouret, et al., Point-of-care versus central laboratory measurements of hemoglobin, hematocrit, glucose, bicarbonate and electrolytes: a prospective observational study in critically ill patients, PLoS One 12 (1) (2017) e0169593.

[7] A.L. Rosenbloom, J.I. Malone, Recognition of impending ketoacidosis delayed by ketone reagent strip failure, JAMA 240 (22) (1978) 2462–2464.

[8] C.J. McKinlay, J.M. Alsweiler, J.M. Ansell, N.S. Anstice, J.G. Chase, G.D. Gamble, et al., Neonatal glycemia and neurodevelopmental outcomes at 2 years, N. Engl. J. Med. 373 (16) (2015) 1507–1518.

[9] American Diabetes Association, Glycemic targets: standards of medical care in diabetes—2019, Diabetes Care 42 (Suppl. 1) (2019) S61–S70.

[10] S.A. Chalew, H. Koetter, S. Hoffman, P.A. Levin, A.A. Kowarski, Diagnosis of reactive hypoglycemia: pitfalls in the use of the oral glucose tolerance test, South. Med. J. 79 (3) (1986) 285–287.

[11] R.L. Nelson, Oral glucose tolerance test: indications and limitations, Mayo Clin. Proc. 63 (3) (1988) 263–269.
[12] J.F. Brun, C. Fedou, J. Mercier, Postprandial reactive hypoglycemia, Diabetes Metab. 26 (5) (2000) 337–351.
[13] L.A. Kollee, L.A. Monnens, V. Cecjka, R.M. Wilms, Persistent neonatal hypoglycaemia due to glucagon deficiency, Arch. Dis. Child. 53 (5) (1978) 422–424.
[14] M.A. Chastain, The glucagonoma syndrome: a review of its features and discussion of new perspectives, Am. J. Med. Sci. 321 (5) (2001) 306–320.
[15] R. Sapin, Insulin assays: previously known and new analytical features, Clin. Lab. 49 (3–4) (2003) 113–121.
[16] A.A. Lauritano, R.S. Clements Jr, D. Bell, Insulin antibodies in non-insulin-dependent diabetes mellitus: effect of treatment with semisynthetic human insulin, Clin. Ther. 11 (2) (1989) 268–277.
[17] N. Hattori, M.R. Duhita, A. Mukai, M. Matsueda, A. Shimatsu, Development of insulin antibodies and changes in titers over a long-term period in patients with type 2 diabetes, Clin. Chim. Acta 433 (2014) 135–138.
[18] R. Wredling, P.E. Lins, U. Adamson, Prevalence of anti-insulin antibodies and its relation to severe hypoglycaemia in insulin-treated diabetic patients, Scand. J. Clin. Lab. Invest. 50 (5) (1990) 551–557.
[19] B. Mianowska, A. Szadkowska, I. Pietrzak, A. Zmysłowska, O. Wegner, J. Tomczonek, et al., Immunogenicity of different brands of human insulin and rapid-acting insulin analogs in insulin-naïve children with type 1 diabetes, Pediatr. Diabetes 12 (2) (2011) 78–84.
[20] A. Maiorana, C. Dionisi-Vici, Hyperinsulinemic hypoglycemia: clinical, molecular and therapeutical novelties, J. Inherit. Metab. Dis. 40 (4) (2017) 531–542.
[21] D.J. Wexler, W. Macias-Konstantopoulos, D.G. Forcione, L. Xiong, C.E. Cauley, K.J. Pierce, Case 23–2018: a 36-year-old man with episodes of confusion and hypoglycemia, N. Engl. J. Med. 379 (4) (2018) 376–385.
[22] P.E. Cryer, L. Axelrod, A.B. Grossman, S.R. Heller, V.M. Montori, E.R. Seaquist, et al., Evaluation and management of adult hypoglycemic disorders: an Endocrine Society Clinical Practice Guideline, J. Clin. Endocrinol. Metab. 94 (3) (2009) 709–728.
[23] K.M. Dreijerink, C.J. Lips, H.T. Timmers, Multiple endocrine neoplasia type 1: a chromatin writer's block, J. Intern. Med. 266 (1) (2009) 53–59.
[24] K.M.A. Dreijerink, H.T.M. Timmers, M. Brown, Twenty years of menin: emerging opportunities for restoration of transcriptional regulation in MEN1, Endocr. Relat. Cancer 24 (10) (2017) T135–T145.
[25] R. Tetzschner, K. Nørgaard, A. Ranjan, Effects of alcohol on plasma glucose and prevention of alcohol-induced hypoglycemia in type 1 diabetes-A systematic review with GRADE, Diabetes Metab. Res. Rev. 34 (3) (2018).
[26] A. Ghosh, I. Banerjee, A.A. Morris, Recognition, assessment and management of hypoglycaemia in childhood, Arch. Dis. Child. 101 (6) (2016) 575–580.
[27] M.W. Haymond, A.S. Pagliara, Ketotic hypoglycaemia, Clin. Endocrinol. Metab. 12 (2) (1983) 447–462.
[28] H. Ozen, Glycogen storage diseases: new perspectives, World J. Gastroenterol. 13 (18) (2007) 2541–2553.
[29] D.A. Weinstein, C.E. Correia, A.C. Saunders, J.I. Wolfsdorf, Hepatic glycogen synthase deficiency: an infrequently recognized cause of ketotic hypoglycemia, Mol. Genet. Metab. 87 (4) (2006) 284–288.
[30] R. Froissart, M. Piraud, A.M. Boudjemline, C. Vianey-Saban, F. Petit, A. Hubert-Buron, et al., Glucose-6-phosphatase deficiency, Orphanet J. Rare Dis. 6 (2011) 27.
[31] K.K. Shah, S.D. O'Dell, Effect of dietary interventions in the maintenance of normoglycaemia in glycogen storage disease type 1a: a systematic review and meta-analysis, J. Hum. Nutr. Diet. 26 (4) (2013) 329–339.
[32] <https://genes-r-us.uthscsa.edu/sites/genes-r-us/files/nbsdisorders.pdf> (accessed 27.02.16).
[33] American Diabetes Association, Classification and diagnosis of diabetes: standards of medical care in diabetes—2019, Diabetes Care 42 (Suppl. 1) (2019) S13–S28.
[34] M.A. Atkinson, G.S. Eisenbarth, A.W. Michels, Type 1 diabetes, Lancet 383 (9911) (2014) 69–82.
[35] W.E. Winter, D.A. Schatz, Autoimmune markers in diabetes, Clin. Chem. 57 (2) (2011) 168–175.
[36] J.D. Cooper, D.J. Smyth, A.M. Smiles, V. Plagnol, N.M. Walker, J.E. Allen, et al., Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci, Nat. Genet. 40 (12) (2008) 1399–1401.
[37] P.Z. Zimmet, T. Tuomi, I.R. Mackay, M.J. Rowley, W. Knowles, M. Cohen, et al., Latent autoimmune diabetes mellitus in adults (LADA): the role of antibodies to glutamic acid decarboxylase in diagnosis and prediction of insulin dependency, Diabet. Med. 11 (3) (1994) 299–303.
[38] B.F. Palmer, D.J. Clegg, Electrolyte and acid-base disturbances in patients with diabetes mellitus, N. Engl. J. Med. 373 (6) (2015) 548–559.
[39] A.L. Rosenbloom, Treatment of diabetic ketoacidosis and the risk of cerebral edema, J. Pediatr. 152 (1) (2008) 146–147.
[40] N. Kuppermann, S. Ghetti, J.E. Schunk, M.J. Stoner, A. Rewers, J.K. McManemy, et al., Clinical trial of fluid infusion rates for pediatric diabetic ketoacidosis, N. Engl. J. Med. 378 (24) (2018) 2275–2287.
[41] W.H. Porter, H.H. Yao, D.G. Karounos, Laboratory and clinical evaluation of assays for beta-hydroxybutyrate, Am. J. Clin. Pathol. 107 (3) (1997) 353–358.
[42] A.E. Kitabchi, G.E. Umpierrez, M.B. Murphy, R.A. Kreisberg, Hyperglycemic crises in adult patients with diabetes: a consensus statement from the American Diabetes Association, Diabetes Care 29 (12) (2006) 2739–2748.
[43] A.L. Rosenbloom, Hyperglycemic crises and their complications in children, J. Pediatr. Endocrinol. Metab. 20 (1) (2007) 5–18.
[44] R. Murphy, S. Ellard, A.T. Hattersley, Clinical implications of a molecular genetic classification of monogenic beta-cell diabetes, Nat. Clin. Pract. Endocrinol. Metab. 4 (4) (2008) 200–213.
[45] L. Aguilar-Bryan, J. Bryan, Neonatal diabetes mellitus, Endocr. Rev. 29 (3) (2008) 265–291.
[46] C.A. Crowther, J.E. Hiller, J.R. Moss, A.J. McPhee, W.S. Jeffries, J.S. Robinson, et al., Effect of treatment of gestational diabetes mellitus on pregnancy outcomes, N. Engl. J. Med. 352 (24) (2005) 2477–2486.
[47] M.B. Landon, et al., A multicenter, randomized trial of treatment for mild gestational diabetes, N. Engl. J. Med. 361 (2009) 1339–1348.

[48] S. Cuhadar, A. Atay, M. Koseoglu, A. Dirican, A. Hur, Stability studies of common biochemical analytes in serum separator tubes with or without gel barrier subjected to various storage conditions, Biochem. Med. (Zagreb.) 22 (2) (2012) 202–214.
[49] W.C. Knowler, E. Barrett-Connor, S.E. Fowler, R.F. Hamman, J. M. Lachin, E.A. Walker, et al., Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin, N. Engl. J. Med. 346 (6) (2002) 393–403.
[50] R.S. Sherwin, R.M. Anderson, J.B. Buse, M.H. Chin, D. Eddy, J. Fradkin, et al., Prevention or delay of type 2 diabetes, Diabetes Care 27 (Suppl. 1) (2004) S47–S54.
[51] American Diabetes Association, Screening for type 2 diabetes, Diabetes Care 27 (Suppl. 1) (2004) S11–S14.
[52] J. Galhardo, J. Shield, The role of haemoglobin A1c in screening obese children and adolescents for glucose intolerance and type 2 diabetes, Acta Med. Port. 28 (3) (2015) 307–315.
[53] S.I. Harashima, A. Nishimura, K. Ikeda, Y. Wang, Y. Liu, N. Inagaki, Once daily self-monitoring of blood glucose (SMBG) improves glycemic control in oral hypoglycemic agents (OHA)-treated diabetes: SMBG-OHA follow-up study, J. Diabetes Sci. Technol. (2015). pii: 1932296815608869.
[54] <https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm611475.htm> (accessed 12.02.18).
[55] L. Jovanovic-Peterson, C.M. Peterson, G.F. Reed, B.E. Metzger, J.L. Mills, R.H. Knopp, et al., Maternal postprandial glucose levels and infant birth weight: the diabetes in early pregnancy study. The National Institute of Child Health and Human Development--Diabetes in early pregnancy study, Am. J. Obstet. Gynecol. 164 (1 Pt 1) (1991) 103–111.
[56] American Diabetes Association., Glycemic targets: standards of medical care in diabetes 2019. Diabetes Care 2019;42(Suppl. 1): S61–S70.
[57] G. van den Berghe, P. Wouters, F. Weekers, C. Verwaest, F. Bruyninckx, M. Schetz, et al., Intensive insulin therapy in the critically ill patients, N. Engl. J. Med. 345 (19) (2001) 1359–1367.
[58] D.E. Griesdale, R.J. de Souza, R.M. van Dam, D.K. Heyland, D.J. Cook, A. Malhotra, et al., Intensive insulin therapy and mortality among critically ill patients: a meta-analysis including NICE-SUGAR study data, CMAJ 180 (8) (2009) 821–827.
[59] American Diabetes Association., Diabetes care in the hospital. Diabetes Care 2016;39(Suppl. 1):S99–S104.
[60] The Diabetes Control and Complications Trial Research Group, The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus, N. Engl. J. Med. 329 (14) (1993) 977–986.
[61] UK Prospective Diabetes Study (UKPDS) Group, Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33), Lancet 352 (9131) (1998) 837–853.
[62] J.T. Jong Taek Kim, W.E. Winter, H. Luo, D.H.K. Chui, N.S. Harris, Interference of hemoglobin A1c due to hemoglobin Franklin Park, J. Appl. Lab. Med. (2018). Available from: https://doi.org/10.1373/jalm.2018.026781.
[63] W.E. Winter, Does availability of hemoglobin A1c results at the point-of-care make a difference in the clinical care or outcome of the patient with diabetes? Point Care Test. J. 5 (2006) 58–61.
[64] M.E. Lacy, G.A. Wellenius, A.E. Sumner, A. Correa, M.R. Carnethon, R.I. Liem, et al., Association of sickle cell trait with hemoglobin A1c in African Americans, JAMA 317 (5) (2017) 507–515.
[65] J. Ahmad, D. Rafat, HbA1c and iron deficiency: a review, Diabetes Metab. Syndr. 7 (2) (2013) 118–122.
[66] R.C. Sundaram, N. Selvaraj, G. Vijayan, et al., Increased plasma malondialdehyde and fructosamine in iron deficiency anemia: effect of treatment, Biomed. Pharmacother. 61 (2007) 682–685.
[67] R.R. Little, D.B. Sacks, HbA1c: how do we measure it and what does it mean? Curr. Opin. Endocrinol. Diabetes Obes. 16 (2) (2009) 113–118.
[68] W.G. John, A. Mosca, C. Weykamp, I. Goodall, HbA(1c) standardisation: history, science and politics, Clin. Biochem. Rev. 28 (4) (2007) 163–168.
[69] W. Hoelzel, C. Weykamp, J.O. Jeppsson, K. Miedema, J.R. Barr, I. Goodall, et al., IFCC reference system for measurement of hemoglobin A1c in human blood and the national standardization schemes in the United States, Japan, and Sweden: a method-comparison study, Clin. Chem. 50 (1) (2004) 166–174.
[70] The DCCT Research Group, Diabetes control and complications trial (DCCT): results of feasibility study, Diabetes Care 10 (1) (1987) 1–19.
[71] D.M. Nathan, J. Kuenen, R. Borg, H. Zheng, D. Schoenfeld, R.J. Heine, et al., Translating the A1C assay into estimated average glucose values, Diabetes Care 31 (8) (2008) 1473–1478.
[72] D.B. Sacks, The diagnosis of diabetes is changing: how implementation of hemoglobin A1c will impact clinical laboratories, Clin. Chem. 55 (9) (2009) 1612–1614.
[73] P.A. Gottlieb, S. Quinlan, H. Krause-Steinrauf, C.J. Greenbaum, D.M. Wilson, H. Rodriguez, et al., Failure to preserve beta-cell function with mycophenolate mofetil and daclizumab combined therapy in patients with new- onset type 1 diabetes, Diabetes Care 33 (4) (2010) 826–832.
[74] C.A. Beam, C. MacCallum, K.C. Herold, D.K. Wherrett, J. Palmer, J. Ludvigsson, et al., GAD vaccine reduces insulin loss in recently diagnosed type 1 diabetes: findings from a Bayesian meta-analysis, Diabetologia 60 (1) (2017) 43–49.
[75] Writing Committee for the Type 1 Diabetes TrialNet Oral Insulin Study Group, J.P. Krischer, D.A. Schatz, B. Bundy, J.S. Skyler, C. J. Greenbaum, Effect of oral insulin on prevention of diabetes in relatives of patients with type 1 diabetes: a randomized clinical trial, JAMA 318 (19) (2017) 1891–1902.
[76] A. Moran, B. Bundy, D.J. Becker, L.A. DiMeglio, S.E. Gitelman, R. Goland, et al., Interleukin-1 antagonism in type 1 diabetes of recent onset: two multicentre, randomised, double-blind, placebo-controlled trials, Lancet 381 (9881) (2013) 1905–1915.
[77] M.D. Pescovitz, C.J. Greenbaum, H. Krause-Steinrauf, D.J. Becker, S.E. Gitelman, R. Goland, et al., Rituximab, B-lymphocyte depletion, and preservation of beta-cell function, N. Engl. J. Med. 361 (22) (2009) 2143–2152.
[78] T. Orban, B. Bundy, D.J. Becker, L.A. DiMeglio, S.E. Gitelman, R. Goland, et al., Co-stimulation modulation with abatacept in patients with recent-onset type 1 diabetes: a randomised, double-blind, placebo-controlled trial, Lancet 378 (9789) (2011) 412–419.
[79] G. Hale , P. Rebello, I. Al Bakir, E. Bolam, P. Wiczling, W.J. Jusko, E. Vandemeulebroucke, B. Keymeulen, C. Mathieu, A.G. Ziegler, L. Chatenoud and H. Waldmann, Pharmacokinetics and antibody responses to the CD3 antibody otelixizumab used in the treatment of type 1 diabetes, *J. Clin. Pharmacol.* **50** (11), 2010, 1238–1248.

[80] P. Ambery, T.W. Donner, N. Biswas, J. Donaldson, J. Parkin, C.M. Dayan, Efficacy and safety of low-dose otelixizumab anti-CD3 monoclonal antibody in preserving C-peptide secretion in adolescent type 1 diabetes: DEFEND-2, a randomized, placebo-controlled, double-blind, multi-centre study, Diabet. Med. 31 (4) (2014) 399–402.
[81] R. Aronson, P.A. Gottlieb, J.S. Christiansen, T.W. Donner, E. Bosi, B.W. Bode, et al., Low-dose otelixizumab anti-CD3 monoclonal antibody DEFEND-1 study: results of the randomized phase III study in recent-onset human type 1 diabetes, Diabetes Care 37 (10) (2014) 2746–2754.
[82] K.C. Herold, S. Gitelman, C. Greenbaum, J. Puck, W. Hagopian, P. Gottlieb, et al., Treatment of patients with new onset Type 1 diabetes with a single course of anti-CD3 mAb Teplizumab preserves insulin production for up to 5 years, Clin. Immunol. 132 (2) (2009) 166–173.
[83] K.C. Herold, S.E. Gitelman, M.R. Ehlers, P.A. Gottlieb, C.J. Greenbaum, W. Hagopian, et al., Teplizumab (anti-CD3 mAb) treatment preserves C-peptide responses in patients with new-onset type 1 diabetes in a randomized controlled trial: metabolic and immunologic features at baseline identify a subgroup of responders, Diabetes 62 (11) (2013) 3766–3774.
[84] N. Sherry, W. Hagopian, J. Ludvigsson, S.M. Jain, J. Wahlen, R.J. Ferry Jr, et al., Teplizumab for treatment of type 1 diabetes (Protégé study): 1-year results from a randomised, placebo-controlled trial, Lancet 378 (9790) (2011) 487–497.
[85] W. Hagopian, R.J. Ferry Jr, N. Sherry, D. Carlin, E. Bonvini, S. Johnson, et al., Teplizumab preserves C-peptide in recent-onset type 1 diabetes: two-year results from the randomized, placebo-controlled Protégé trial, Diabetes 62 (11) (2013) 3901–3908.
[86] K.C. Herold, B.N. Bundy, S.A. Long, J.A. Bluestone, L.A. DiMeglio, M.J. Dufort, et al., N. Engl. J. Med. 381 (7) (2019) 603–613.
[87] U. Zafar, S. Khaliq, H.U. Ahmad, S. Manzoor, K.P. Lone, Metabolic syndrome: an update on diagnostic criteria, pathogenesis, and genetic links, Hormones (Athens) 17 (3) (2018) 299–313.
[88] D.J. Timson, The molecular basis of galactosemia - past, present and future, Gene (2015).

Self-assessment questions

1. A comatose patient who is on IV fluids who is taking nothing by mouth is in septic shock. A finger-stick capillary glucose is measured and is reported at <60 mg/dL (the lower limit of detection of the analyzer is 60 mg/dL and the reference interval for FPG is 70–99 mg/dL). What action should the clinical team take?
   a. Administer IV glucose (IV = intravenous injection).
   b. Administer IV insulin (IV = intravenous injection).
   c. Administer IV glucagon (IV = intravenous injection).
   d. Send a venipuncture sample to the lab for glucose measurement.
   e. Take no action at this time. No intervention is required.
2. A sample arrives in the core lab from an outside draw station 18 hours after being obtained. The patient is a 60-year-old healthy man who was being screened for diabetes mellitus as part of routine health care. The serum glucose in the red top tube is 40 mg/dL (reference interval for fasting serum glucose: 70–99 mg/dL). What is the most likely cause for this result?
   a. The tube was an serum separator tube and was spun prior to transport.
   b. The patient contracted his forearm repeatedly during the phlebotomy.
   c. Hyperinsulinism is likely present.
   d. The tube is likely a sodium fluoride tube and is not a red top tube.
   e. Delayed analysis is the cause of the low serum glucose result.
3. An adult woman with a history compatible with hypoglycemia undergoes an inpatient fast. Her glucose declines to 35 mg/dL at 8 hours into the fast and she feels very tired. Glucagon is administered, and 45 minutes later, her plasma glucose concentration rises to 100 mg/dL. After glucagon, the patient feels awake. What is the most likely diagnosis?
   a. Liver failure
   b. Hyperinsulinism
   c. Gluconeogenic defect
   d. Cortisol deficiency
   e. GH excess
4. In an adult woman, hypoglycemia is confirmed during a 72-hour fast. At the time of hypoglycemia, insulin and C-peptide are both elevated above the reference intervals for a normal FPG. What drug ingestion should be excluded before the workup for insulinoma begins in earnest?
   a. Ethanol
   b. Metformin
   c. Sulfonylureas
   d. Statins
   e. Cocaine
5. A 65-year-old lean man presents with polyuria and polydipsia. He has mild persistent hyperglycemia (FPG = 130 mg/dL; reference interval = 70–99 mg/dL) and an HbA1c of 6.9% (reference interval $\leq 6.5\%$). He is normotensive and has a normal lipid profile. What testing should next be ordered to determine the cause of his diabetes?
   a. Islet autoantibodies
   b. Insulin and C-peptide
   c. Cortisol and growth hormone
   d. Glucagon and epinephrine
   e. HLA typing
6. A 1-month-old infant develops dehydration and polyuria. The plasma glucose is found to be 300 mg/dL (reference interval for FPG: 70–99 mg/dL). What is the most likely etiology of the baby's problems?
   a. Glucagonoma
   b. Autoimmune beta cell destruction
   c. An inborn error in the beta cell

d. Pheochromocytoma
e. Insulin resistance

7. A woman of reproductive age reports to her family practitioner than she missed 1 menstrual period. In the office, her pregnancy test is positive. The patient's BMI is 33 kg/M^2, her blood pressure is elevated, and she has acanthosis nigricans. Her FPG is 145 mg/dL (reference interval = 70−99 mg/dL) and her HbA1c is 7.6% (reference interval $\leq$ 6.5%). What is the correct diagnosis?
a. T1DM mellitus
b. Type 2 diabetes mellitus
c. Impaired glucose tolerance
d. GDM
e. Normoglycemia

8. Comparing the diagnostic criteria for diabetes mellitus in nonpregnant adults and children, both do these criteria differ?
a. There are different definitions of fasting hyperglycemia.
b. There are different definitions for hyperglycemia at the 2-hour time point during an OGTT.
c. There are different definitions for hyperglycemia at the zero time point during an OGTT.
d. There are different definitions for an elevated HbA1c.
e. There are no differences in the definitions of hyperglycemia.

Answers

1. d
2. e
3. b
4. c
5. a
6. c
7. b
8. e

Chapter 35

Laboratory evaluation of kidney function

W. Greg Miller[1] and Lesley A. Inker[2]

[1]*Department of Pathology, Virginia Commonwealth University, Richmond, VA, United States,* [2]*Department of Medicine, Division of Nephrology, Tufts Medical Center, Boston, MA, United States*

Learning objectives

After reviewing this chapter, the reader should be able to:

- Discuss the physiology of kidney function.
- Explain the categories of chronic kidney disease.
- Measure and estimate glomerular filtration rate.
- Explain laboratory measurement of biomarkers of kidney function.
- Discuss limitations of biomarkers to assess kidney function.

Physiology of kidney function

The overall function of the kidney is to maintain a relatively constant extracellular environment necessary for normal cellular function and to eliminate metabolic waste products [1]. The kidney monitors the internal physiological environment and achieves a constant extracellular environment by secretion and excretion of products derived from dietary intake (e.g., water, hydrogen ion, and electrolytes) and of waste products derived from metabolism (e.g., urea, creatinine, and uric acid). The kidney also secretes hormones (e.g., renin, erythropoietin, and vitamin D) important for metabolic processes. Removal of metabolic waste products and toxins is called "clearance" and is achieved by filtration and secretion. Most of the kidney's functions can be achieved with only one kidney or low function in both kidneys, and indeed, because of adaptive processes, there may be minimal signs or symptoms of kidney damage until a severe reduction [usually defined as a glomerular filtration rate (GFR) of <30 mL/min/1.73 m^2] in kidney function occurs. The term uremia refers to the clinical signs and symptoms that occur with kidney failure due to accumulation of metabolic waste products. The specific products are not known, but they are generally thought to occur at GFR of <15 mL/min/1.73 m^2 or lower. Some consider the blood urea nitrogen (BUN) level as a marker of uremia, but even BUN concentrations of >100 mg/dL may not have apparent clinical effects in many clinical circumstances. In addition, BUN concentrations do not correlate with other retained nitrogenous waste products. Reduction in dietary protein intake does reduce the BUN without necessarily changing symptoms.

The structure of the kidney is designed to accomplish filtration, reabsorption, and secretion functions (Fig. 35.1). There are approximately 1 million individual anatomic units in each kidney called nephrons. Each nephron contains a glomerulus that is connected to a twisted tubule section that drains into a collecting duct and accompanying blood supply. The collecting ducts gradually develop into the urine collecting system at the renal pelvis. The renal pelvis of each kidney drains into a single ureter that carries the urine to the bladder.

The glomerulus receives blood from systemic circulation via the afferent arteriole. After filtration at the glomerulus, blood enters a capillary system called the vasa recta that perfuses the tubule segments and then returns to the body via the efferent arteriole. The glomerulus is composed of an interconnected series of capillary loops enclosed within Bowman's capsule. The role of the glomerulus is to filter water and small molecules, but retain larger molecules such as proteins and blood cells. The glomerular basement membrane is selective based on size and charge of a molecule. The pressure in the glomerulus is an important component of filtration. Surrounding mesangial cells assist to control glomerular surface area and filtration fraction.

The filtered fluids pass from Bowman's capsule into the tubules. Specific tubular segments perform excretion of water, electrolyte and hydrogen ion handling, and hormone production (renin, 1,25-dihydroxy vitamin D, and erythropoietin). The vasa recta bring the blood supply to the tubules and allow for continuous removal of waste products and toxins via tubular secretion into the filtration fluid, and reclaiming of needed molecules and ions via tubular reabsorption from the filtration fluid. The proximal tubule reabsorbs 70% of the filtered sodium and chloride and most of the glucose, amino acids, bicarbonate, potassium, phosphate, and sulfate. The homeostasis of water is

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00035-1

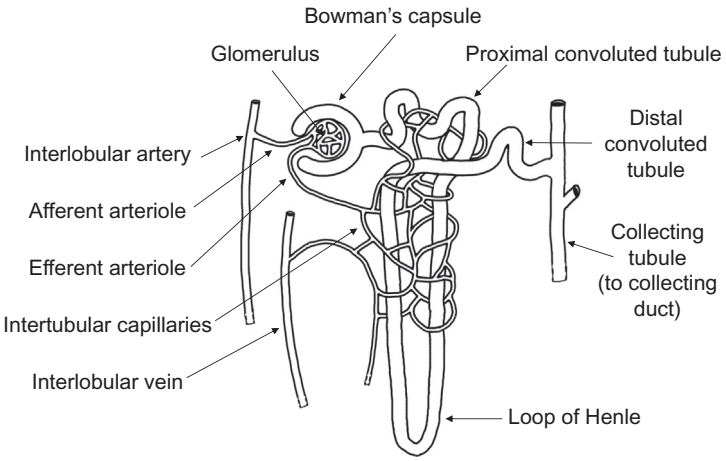


FIGURE 35.1 Anatomic structure of a kidney nephron and collecting duct.

regulated by the varying permeability along the segments of the loop of Henle. The thin descending limb is permeable and the ascending limb is impermeable to water reabsorption, with passive diffusion of NaCl along the thin ascending limb and active reabsorption along the thick ascending limb. Thus a highly dilute urine enters the distal tubule. In the presence of antidiuretic hormone (ADH), aquaporins are inserted in the distal collecting duct principle cells, allowing movement of water from lumen to the interstitium and reabsorption of water. Thus the concentration of urine is based on the presence or absence of ADH. (Water and electrolyte balance is covered in Chapter 37) Renin is produced by the nephrons in response to blood volume and stimulates production of aldosterone by the pituitary. Aldosterone acts on the distal tubule to manage final water and sodium retention, and exchange of potassium and hydrogen ions (see Chapter 41, Adrenal function).

The kidney receives 20% of the cardiac output, and maintenance of systemic blood pressure is a major function of the kidney. The kidneys have the ability to maintain glomerular filtration despite wide fluctuations in systemic hemodynamics through a process called adaptation. Adaptation occurs via changes in hemodynamics, which are controlled by the renin−angiotensin−aldosterone system and other molecules such as prostaglandins. Prostaglandin-dependent vasodilation of the afferent arteriole preserves glomerular perfusion despite falls in kidney blood flow. Conversely, vasoconstriction at times of high systemic pressure protects the glomerular capillary from potential damage resulting from high blood pressure.

Chronic kidney disease

Definition

Chronic kidney disease (CKD) is defined by the Kidney Disease: Improving Global Outcomes (KDIGO) Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease as abnormalities of kidney structure or function, present for 3 or more months, with implications for health [2]. The primary biomarker for abnormal kidney function is GFR of <60 mL/min/1.73 m^2 and for kidney damage is urine albumin-to-creatinine ratio (ACR) of >30-mg albumin per gram of creatinine. In addition, urine sediment, electrolyte and tubular disorders, biopsy, and imaging may indicate kidney abnormality. Kidney damage is usually ascertained by biomarkers without a kidney biopsy.

CKD is recognized as a major public health problem, given the increasing prevalence of kidney failure with poor outcomes and high costs over the past decades. Kidney failure is the "end stage" of CKD and abbreviated end-stage renal disease (ESRD). Common risk factors for CKD include hypertension, diabetes, obesity, cardiovascular disease (CVD), and older age [2]. The prevalence of overall CKD is estimated at 12% of the U.S. population [3]. The number of persons with kidney failure who are treated with dialysis and transplantation in the United States is projected to increase from 470,000 in 2004 to over 2.2 million in 2030, and was 661,468 in 2013 [4].

The GFR is usually accepted as the best overall index of kidney function in health and disease. The normal GFR varies according to age, sex, and body size. Normal GFR in young adults is approximately 120−130 mL/min/1.73 m^2. A GFR of <60 mL/min/1.73 m^2 represents the loss of half or more of normal adult kidney function and is associated with an increased prevalence of systemic complications. GFR declines with age by approximately 1 mL/min/1.73 m^2/year after the third decade. More than 25% of individuals aged ≥ 70 years have GFR of <60 mL/min/1.73 m^2. The etiology of the reduced GFR in the elderly is not known. Some suggest that this may be due to normal aging, although there are many older adults who do not have decreased GFR, suggesting the decreased GFR may be secondary to subclinical forms of systemic diseases that cause kidney damage.

Categories of chronic kidney disease

The categories of CKD are shown in Tables 35.1 and 35.2 as described in the KDIGO guideline. The categories are defined using the two biomarkers GFR and ACR. The categories of CKD can be represented in a two-dimensional chart (Fig. 35.2), showing the combination of GFR and ACR values with different shades in the grid, representing the severity of the disease associated with different values for the biomarkers [2]. The adverse outcomes of CKD are not restricted to kidney failure but also include the

TABLE 35.1 Categories of chronic kidney disease by the level of glomerular filtration rate.

Category	GFR levels (ml/min/1.73 m^2)	Terms	Clinical action plan
G1[a]	>90	Normal or high	• Diagnose and treat the cause • Treat comorbid conditions • Evaluate for CKD risk factors • Start measures to slow CKD progression • Start measures to reduce CVD risk
G2[a]	60–89	Mildly decreased[b]	• Estimate progression
G3a	45–59	Mildly to moderately decreased	• Adjust medication dosages as indicated • Evaluate and treat complications
G3b	30–44	Moderately to severely decreased	
G4	15–29	Severely decreased	• Prepare for kidney replacement therapy (transplantation and/or dialysis)
G5	<15	Kidney failure (add D if treated by dialysis)	• Start kidney replacement therapy (if uremia present)

Notes: GFR in mL/min/1.73 m^2 may be converted to mL/s/1.73 m^2 by multiplying by 0.01667. *GFR*, Glomerular filtration rate; *CKD*, chronic kidney disease; *CVD*, cardiovascular diseases.

[a]*GFR stages G1 or G2 without markers of kidney damage do not fulfil the criteria for CKD.*
[b]*Relative to young adult level.*
Source: Reprinted with permission from Chapter 94 in the Oxford Textbook of Clinical Nephrology (fourth ed.), N. Turner, N. Lameire, D.J. Goldsmith, C.G. Winearls, J. Himmelfarb, G. Remuzzi (Eds.), Oxford University Press, 2015.

TABLE 35.2 Categories of chronic kidney disease by the level of albuminuria.

Category	AER	Approximately equivalent ACR		Terms	Clinical action plan
	(mg/d)	(mg/mmol)	(mg/g)		
A1	<30	<3	<30	Normal to mildly increased[a]	• Diagnose and treat the cause • Treat comorbid conditions • Evaluate for CKD risk factors • Start measures to slow CKD progression • Start measures to reduce CVD risk
A2	30–299	3–30	30–299	Moderately increased	• Treatment with renin–angiotensin system blockers and lower blood pressure goal if hypertensive
A3	≥300	≥30	≥300	Severely increased	• Treat nephritic or nephrotic syndrome (if present)

ACR, Albumin-to-creatinine ratio; *AER*, albumin excretion rate; *CKD*, chronic kidney disease; *CVD*, cardiovascular diseases.

[a]*Relative to young adult level.*
Source: Reprinted with permission from Chapter 94 in the Oxford Textbook of Clinical Nephrology (fourth ed.), N. Turner, N. Lameire, D.J. Goldsmith, C.G. Winearls, J. Himmelfarb, G. Remuzzi (Eds.), Oxford University Press, 2015.

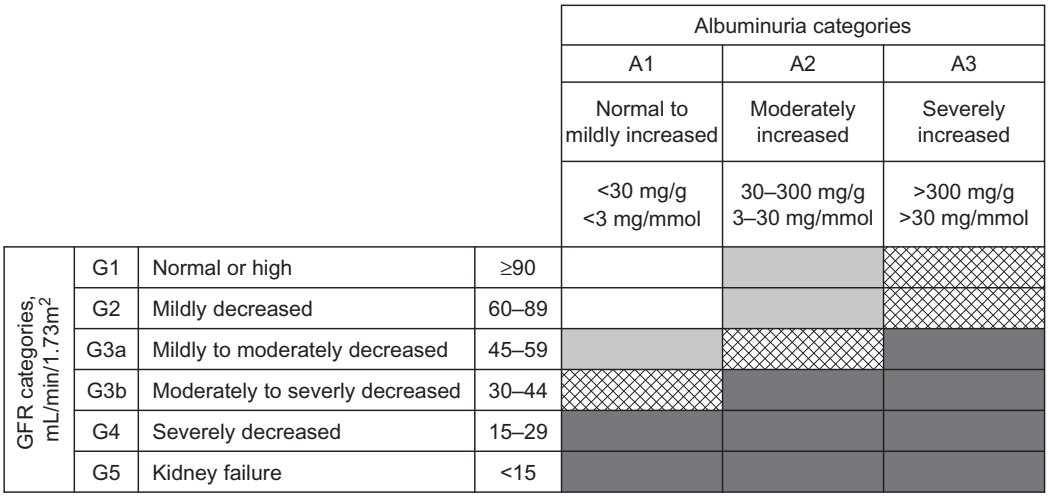


FIGURE 35.2 Prognosis of chronic kidney disease by glomerular filtration rate and albuminuria categories with different shades in the grid representing the severity of the disease associated with different values for the biomarkers (clear nondiseased; light grey, crosshatch, and dark grey indicate increasing severity). *Adapted from Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group, KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease, Kidney Int. 3 (2013) 1–150. (https://kdigo.org/guidelines/ckd-evaluation-and-management/, accessed 06.03.2020).*

complications of decreased kidney function, such as hypertension, anemia, malnutrition, bone and mineral disorders, neuropathy, and increased risk of CVD. In addition, there is evidence for associations with cognitive impairment, infections, frailty, and decreased functional abilities. Risk for CVD, ESRD, and mortality start at levels of GFR of >60 mL/min/1.73 m^2 in most populations and increase linearly with decreasing GFR [5]. Similarly, for albuminuria, the risk for these outcomes appears to increase as a continuous function from concentrations even less than 30-mg albumin/g creatinine. For the other complications mentioned above, risk appears to correlate more with GFR than ACR, with increasing risk beginning at levels of GFR of <60 ml/min/1.73 m^2. Risk for CVD, ESRD, and mortality increases with lower GFR and higher albuminuria, and most recent data suggest a multiplicative relationship between the two [2,5].

Public health efforts

In 2002, the Kidney Disease Outcomes Quality Improvement (KDOQI) guidelines for identification and treatment of CKD [3] reported that millions of people with category 3 CKD were not being identified and consequently were not being treated. Treatment at this category was able to substantially slow progression of disease with reduction in complications, morbidity, and mortality. It was recognized that, although creatinine was commonly measured as part of many clinical encounters, physicians had difficulty relating small increases in concentration with the appropriate decrease in kidney function (see "Limitations of serum creatinine interpretation" section). The KDOQI guidelines recommended that estimated glomerular filtration rate (eGFR) calculated from serum (or plasma) creatinine be reported with the creatinine value to provide physicians with a parameter that was more easily understood and interpreted.

The National Kidney Disease Education Program (NKDEP) was formed by the National Institutes for Diabetes, Digestive and Kidney Diseases at the National Institutes of Health (United States) to educate high-risk patients, physicians, and clinical laboratories about CKD and how to identify patients for treatment. In 2006, the NKDEP recommended that clinical laboratories in the United States report eGFR when serum creatinine is reported [6]. Estimated GFR reporting practices have been reviewed and >90% of clinical laboratories in the United States now follow this practice [7]. Most laboratory and nephrology professional associations around the world have endorsed reporting eGFR. In most circumstances, eGFR is sufficient for clinical decision-making. However, as described below, for some patients, the error in eGFR can be substantial and a measured GFR is needed.

More recently, the National Kidney Foundation (United States) and the American Society for Clinical Pathology have recommended that laboratories offer a "Kidney Profile" that includes serum creatinine, calculated eGFR using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation (see later section), and urine ACR [8]. The "Kidney Profile" follows evidence-based clinical practice guidelines to ensure physicians order the appropriate group of tests to more effectively

assess kidney function for patients, such as diabetics and hypertensives, who are at increased risk for CKD [9].

Measurement of glomerular filtration rate

Clearance methods

Clearance of a substance from the plasma as it passes through the glomerulus is used to measure the GFR. The ideal substance is a low-molecular-weight water-soluble molecule that freely crosses the glomerular membrane, is not bound to plasma proteins, is not reabsorbed nor secreted nor metabolized by kidney tubules, and is not eliminated by any nonkidney process.

Clearance is measured by elimination of an exogenous filtration marker from the blood. For such a substance, the amount that appears in the urine will equal the amount that crosses the glomerulus, and the rate of plasma flow through the glomerulus (the GFR) can be calculated as:

$$\begin{aligned}\text{GFR} = {} & (\text{concentration in urine}) \times (\text{volume of urine})/ \\ & (\text{concentration in plasma}) \\ & \times (\text{time of urine collection})(\text{units: mL/min})\end{aligned} \tag{35.1}$$

where the urine and plasma concentrations are in the same units (e.g., milligrams per deciliter or micromoles per liter), the volume of urine is in milliliter, and the time of urine collection is in minutes. Urine clearance measurements require a timed urine collection to average influence of diurnal variation in rates of excretion with hydration and physical activity. The timed urine collection is generally 24 h but can be shorter, especially when it can be performed under observation or in acute care settings. The primary disadvantage of urinary clearance measurements is the need for urinary collection, which is especially a concern in populations with incomplete bladder emptying, such as children and the elderly.

There is an increasing interest in measurement of clearance from plasma, because it avoids the need for a timed urine collection. GFR is calculated from plasma clearance of the marker after a bolus intravenous injection of an exogenous filtration marker. The GFR is calculated from the dose and the area underneath the disappearance curve. The decline in plasma concentration is secondary to the immediate disappearance of the marker from the plasma into its volume of distribution and a slower renal clearance. Plasma clearance is best estimated using a two-compartment model that requires blood sampling early (usually two or three time points until 60 min) to estimate the fast curve for volume of distribution and late (one to three time points from 120 to 300 min) to estimate the slow curve for renal clearance. For patients with very low levels of kidney function, a very late time point (e.g., at 24 h) is necessary to avoid overestimation of the GFR.

Kidney size is approximately proportional to body size. To enable standardized interpretation of GFR, the values obtained from plasma or urinary clearance methods are typically adjusted to a body surface area (BSA) of 1.73 m^2. BSA is calculated as [10]:

$$\begin{aligned}\text{BSA} = {} & \text{weight(kg)}^{0.425} \times \text{height(cm)}^{0.725} \\ & \times 7.1 \times 10^{-3}\left(\text{units:}m^2\right).\end{aligned} \tag{35.2}$$

The GFR corrected to 1.73-m^2 BSA is:

$$\text{GFR(adjusted)} = \text{GFR} \times 1.73/\text{BSA}\left(\text{units:mL/min/1.73}\,m^2\right). \tag{35.3}$$

Exogenous substances to measure glomerular filtration rate

Inulin has the properties of an ideal substance to measure GFR and has been the exogenous substance used as the gold standard. Classically, inulin is given via intravenous infusion using bladder catheterization to ensure complete bladder emptying. However this method is invasive, and inulin is difficult to maintain in solution and to measure accurately and as such is not often used in clinical practice. Other substances and methods have been used to measure GFR and are listed in Table 35.3. The most common markers used in the United States are iohexol and iothalamate. For all markers currently used, there have been concerns expressed regarding their ideal characteristics for renal handling, such that inulin remains the true gold standard. Nevertheless, with accurate assay calibration and quality control techniques, these alternative methods and markers can be used as reference standards in clinical practice and research studies [11].

Endogenous substances to measure glomerular filtration rate

Clearance of endogenous substances has been commonly used to measure GFR. An ideal endogenous filtration marker shares the same properties as an ideal exogenous marker; however, in addition, it is generated consistently by the body to maintain a constant plasma concentration. No endogenous substances have such ideal properties, and all endogenous filtration markers have non-GFR determinants of their plasma concentrations due to variation in generation of the biomarker within and between individuals, as well as excretion from the plasma space not due to glomerular filtration. The most common endogenous markers to estimate GFR are creatinine and cystatin C.

TABLE 35.3 Procedures used to measure glomerular filtration rate.

Marker	Method of administration	Comments
Inulin	Continuous IV	Gold standard; difficult to handle and measure accurately
Iothalamate	Bolus IV or subcutaneous	Can be administered as a radioactive compound with ^{125}I as the tracer or as a nonradioactive compound with assays using HPLC methods. In radioactive form, potential problem of thyroid uptake of ^{125}I. Iothalamate is secreted, leading to overestimation of GFR
99mTc-DTPA	Bolus IV	Dissociation of 99mTc leads to plasma protein binding and underestimation of GFR
51Cr-EDTA	Bolus IV	10% lower clearance than inulin
Iohexol	Bolus IV	Low incidence of adverse effects. Comparable with inulin. Assay by HPLC and IDMS.

GFR, Glomerular filtration rate; *HPLC,* high-performance liquid chromatography; *IDMS,* isotope dilution mass spectrometry.

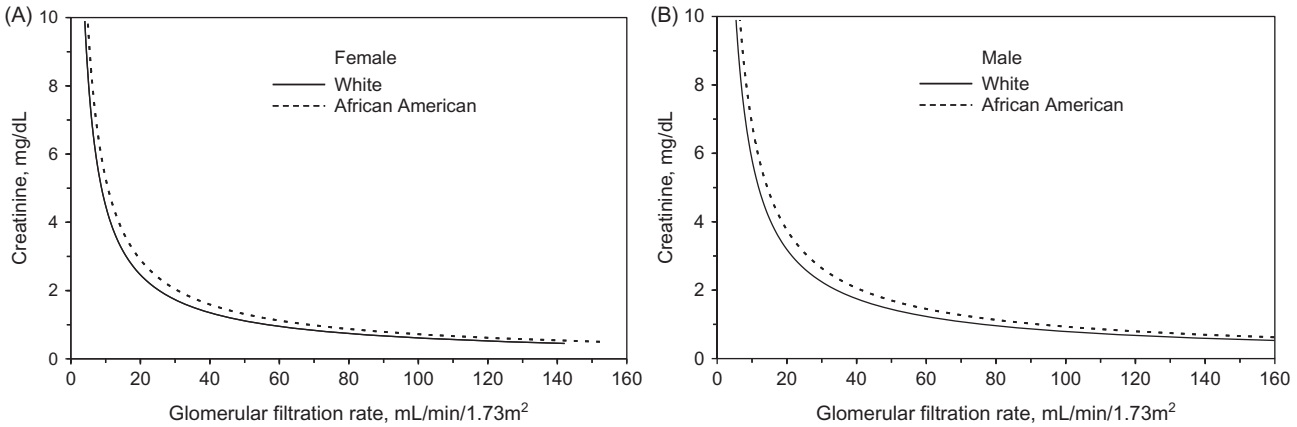


FIGURE 35.3 Inverse nonlinear relationship between serum/plasma creatinine and glomerular filtration rate. The relationship was estimated using the modification of diet in renal disease study equation for a male and a female 60-year-old individual.

Measured clearance of endogenous substances is more convenient than measured clearance of exogenous filtration markers and therefore provides a reasonably accurate method to estimate the GFR. However, because of the need for timed urine collections, these measurements are difficult to do correctly and lead to substantial errors in results. In addition, the presence of tubular reabsorption and secretion of the marker will lead to underestimate or overestimate of the GFR, respectively.

Creatinine

Physiology of creatinine

Creatinine is produced from breakdown of creatine in muscle metabolism and presents a relatively constant concentration in plasma that enters the glomerulus. Creatinine is freely filtered by the glomerulus and is excreted in the urine. Fig. 35.3 shows the inverse relationship between serum/plasma creatinine and GFR [12]. Note that the relatively small changes in creatinine concentration at earlier stages of kidney function decline (e.g., a GFR change from 120 to 60 mL/min/1.73 m^2) correspond to a serum creatinine increase from 0.8 to 1.4 mg/dL for a 30-year-old white male. However, there is substantial interindividual variation in the generation of creatinine that is primarily related to differences in muscle mass and diet. In addition, creatinine is moderately secreted (7%–10% in healthy individuals) by proximal tubules, and the rate of secretion typically increases as CKD progresses. Sickle cell disease can increase tubular secretion. Consequently, creatinine clearance is numerically larger than GFR, and the positive bias increases as kidney function deteriorates. Some conditions such as acute kidney necrosis can decrease tubular

secretion of creatinine, causing decreased creatinine excretion. Medications and specific diseases can affect secretion; most notably cimetidine and trimethoprim; dolutegravir and cobicistat used in antiretroviral regimens for HIV management; and poly(ADP-ribose) polymerase inhibitors, such as olaparib and rucaparib, and tyrosine kinase inhibitors, such as crizotinib, gefitinib, imatinib, pazopanib, sorafenib, and sunitinib used to treat cancer can block tubular secretion of creatinine. In addition, proximal tubulopathies that develop secondary to medications, such as tenofovir, cisplatin, or heavy metals, can lead to decreased creatinine secretion. Finally, there is some elimination of creatinine via the gut. Gut secretion is difficult to measure but likely increases with worsening kidney disease. A measured GFR should be used when conditions that alter creatinine elimination are present.

Measurement of creatinine

Creatinine can be measured by its reaction with alkaline picric ions (first described by Jaffe in 1886) to form an orange–red complex of unknown chemical structure. More recently, procedures using enzymes have been introduced. The Jaffe reaction is not specific for creatinine. Serum (or plasma) proteins, glucose, ketone bodies, ascorbate, bilirubin, hemoglobin F, other endogenous molecules, and several exogenous drugs also react to produce varying amounts of chromogenic product with picrate ion. These interfering substances have been called "noncreatinine chromogens." Implementations of the Jaffe reaction have been developed that varied the reaction conditions to minimize the influence of some of these potentially interfering substances. For example, some interfering substances react very quickly to form the picrate complex (e.g., acetoacetate) and others more slowly (e.g., albumin and immunoglobulins). Making a kinetic measurement of the reaction between approximately 20 and 60 s optimizes detection of the creatinine reaction period and minimizes the influence of several other substances. The reagent formulation and temperature have an important influence on reactivity of creatinine and other substances in the Jaffe reaction.

Another approach used by several manufacturers is to subtract a constant concentration from the measured value with a Jaffe reaction to correct (called compensate) for the average amount of noncreatinine chromogens in a sample. This approach works well as long as an individual sample has the same amount of noncreatinine chromogens as the average samples used to develop the correction factor. The compensation approach mainly corrects for proteins, not for other substances only found in some samples from diseased patients. Compensated methods have a limitation when used for pediatric or hospitalized patients who frequently have low albumin or immunoglobulin concentrations. In these cases, the correction is too large, causing an erroneously low creatinine concentration. The error can be a large percent change for younger children who may have creatinine concentrations in the range of 0.3–0.6 mg/dL.

Creatinine can also been measured using coupled enzyme reactions in an effort to develop more specific methods. Several reaction schemes have been developed. One of the more common approaches uses creatininase to convert creatinine to creatine, creatinase to convert the creatine to sarcosine and urea, and sarcosine oxidase to produce products including hydrogen peroxide, which is subsequently reduced to form a colored chromogen. Creatininase is specific for creatinine, and endogenous creatine and urea are removed in a prereaction incubation before creatininase is added. Some interfering substances may influence enzymatic methods, primarily by their reactions in the several coupled reaction steps. For example, bilirubin and ascorbate can react with hydrogen peroxide to decrease the amount of chromogen produced. In general, enzymatic methods are affected by fewer interfering substances than Jaffe-based methods; in particular, proteins do not interfere in enzymatic methods, making them well suited to pediatric and hospitalized patient groups.

Much of the older literature investigating interfering substances was not performed with the current implementations of Jaffe and enzymatic methods. A recent report examined the performance of four enzymatic and three Jaffe routine methods compared with an LC-isotope dilution mass spectrometry (IDMS) procedure when measuring individual patient samples selected to contain commonly encountered interfering substances [13]. Both Jaffe and enzymatic methods were influenced by interfering substances to different degrees. However, Jaffe methods were more susceptible to interfering substances particularly for patients with diabetes, CVD, protein abnormalities, and cephalosporin drugs. Enzymatic methods were minimally affected by those substances but were subject to interferences from hemolysis and elevated bilirubin that are easily identified by contemporary measuring systems. In addition, the influence of interfering substances was different for different implementations of a Jaffe or enzymatic procedure. Recent reports suggest that enzymatic methods for creatinine are preferred due to their improved selectivity, better precision, and smaller bias versus reference measurement procedures that are particularly important for pediatric populations [14–16].

Standardized calibration of creatinine measurements

Prior to approximately 2010, calibration of creatinine methods was poorly standardized among different methods and laboratories. Biases were reported to be 30%–40%

among different methods [6]. Consequently, different reference intervals were used by different methods, and uniform application of clinical practice guidelines was difficult. When the KDOQI recommended reporting eGFR with all creatinine results, the Laboratory Working Group of the NKDEP initiated a calibration standardization program for creatinine methods (including serum, plasma and whole blood samples). All creatinine methods from global manufacturers now have standardized calibration traceable to a Joint Committee for Traceability in Laboratory Medicine (JCTLM) listed IDMS reference measurement procedure that itself is calibrated with a primary reference material [SRM 914a crystalline creatinine from the National Institute of Standards and Technology (NIST)] [17]. A creatinine in frozen human serum reference material was developed by NIST (SRM 967) as a trueness control to assist in documenting traceability.

Limitations of serum creatinine interpretation

It has been difficult to establish useful adult reference intervals for creatinine, because a reference population without kidney disease is difficult to identify. Early CKD is not recognized by any clinical symptoms or laboratory biomarkers. Consequently, it is likely that some people in an apparently healthy reference population have some degree of CKD. As mentioned earlier, GFR declines approximately 1 mL/min/1.73 m^2/year after the third decade, which complicates selecting a reference population. A further complication is that, because of the interindividual variation in creatinine generation, reference intervals differ by gender, age, and race to reflect differences in muscle mass. Reference intervals have typically been stratified by age and sex for children and only by sex for adults. It is rare that laboratories stratify by age or race in adults despite changes in muscle mass in these groups. In addition, even stratified reference intervals cannot incorporate other causes for differences in creatinine generation, such as changes in diet or other extremes of muscle loss (e.g., amputation or muscle-wasting diseases). Consequently, reference intervals have not been satisfactory to discriminate normal from moderately diseased patients and physicians have had difficulty in identifying patients in early stages of kidney disease only on the basis of serum creatinine concentrations. Equations have been developed to transform serum creatinine into eGFR, which is more easily understood by physicians to identify patients with early kidney disease who can be helped by therapeutic intervention (see *Estimating Equations*).

Cystatin C

Physiology of cystatin C

Cystatin C is a lysosomal cysteine proteinase (13,000 Da) produced at a constant rate in nucleated cells. It is freely filtered at the glomerulus, not secreted by the tubules, and not eliminated other than by the kidney, and is completely reabsorbed and catabolized by the proximal tubules. Like creatinine, cystatin C concentration has an inverse nonlinear relationship to GFR. Cystatin C does not appear in the urine unless there is tubular damage. Consequently, urinary clearance of cystatin cannot be determined.

Cystatin C concentration in plasma has traditionally been thought to be generated at a constant rate by all nucleated cells. Lack of urinary clearance has made this hypothesis difficult to thoroughly evaluate, but data suggest that there are non-GFR determinants because of interindividual variation in generation of cystatin [18]. Factors that affect the concentration of cystatin C include age, sex, diabetes, inflammation, body habitus, urinary protein, thyroid status, and steroid use. Cystatin C is not dependent on muscle mass and diet and is useful in conditions when creatinine is not suitable.

Measurement of cystatin C

Cystatin C is measured by immunoassay. In-vitro diagnostic device (IVD) manufacturers have used different antibodies directed against different epitopes. An international federation of clinical chemistry and laboratory medicine (IFCC) working group for standardization of cystatin C has developed a certified reference material, ERM DA471/IFCC, that has been available since 2010 from the Joint Research Center of the European Union (JRC, EU). The certified reference material consists of recombinant cystatin C in a human serum pool and is intended for calibration of immunoassays. Some manufacturers have recalibrated their cystatin C methods to be traceable to the ERM DA471/IFCC reference material, and all are expected to achieve standardized calibration traceability. The availability of standardized cystatin C methods has enabled estimating equations for eGFR to be developed from large populations that can be used with any standardized cystatin C method. However, as of this writing, calibration of cystatin C assays is not completely standardized, and different methods give different numeric results [19,20].

Estimating equations

Estimated glomerular filtration rate

Estimating equations combine an endogenous filtration marker (e.g., creatinine or cystatin C) with easily measured surrogates for other non-GFR determinants that influence the serum concentration of the endogenous filtration marker. For both markers, the non-GFR determinants age, sex, race, or body size have been incorporated with creatinine or cystatin C in estimating equations. A summary of common eGFR equations can be found in

TABLE 35.4 Equations for estimation of glomerular filtration rate.

1) Cockcroft–Gault
CrCl (male) = ([140 age] × weight in kilogram)/(Scr × 72) if female, a multiplier of 0.85 is used where Scr is serum creatinine in mg/dL, and age is in years
2) MDRD (four-variable)
GFR (mL/min/1.73 m^2) = 175 × $(S_{cr})^{-1.154}$ × $(Age)^{-0.203}$ × (0.742 if female) × (1.212 if African American) where Scr is serum creatinine in mg/dL, and Age is in years
3) CKD-EPI creatinine
GFR = 141 × $\min(S_{cr}/\kappa, 1)^{\alpha}$ × $\max(S_{cr}/\kappa, 1)^{-1.209}$ × 0.993^{Age} × 1.018 [if female] × 1.159 [if black] where S_{cr} is serum creatinine in mg/dL, κ is 0.7 for females and 0.9 for males, α is −0.329 for females and −0.411 for males, min indicates the minimum of S_{cr}/κ or 1, and max indicates the maximum of S_{cr}/κ or 1, and Age is in years
4) CKD-EPI 2012 cystatin C
eGFR = 133 × $\min(S_{cys}/0.8, 1)^{-0.499}$ × $\max(S_{cys}/0.8, 1)^{-1.328}$ × 0.996^{Age} × 0.932 [if female] where Scys is cystatin C in mg/L, and Age is in years min indicates the minimum of $S_{cys}/0.8$ or 1 max indicates the maximum of $S_{cys}/0.8$ or 1
5) CKD EPI 2012 cystatin C–creatinine
eGFR = 135 × $\min(S_{cr}/\kappa, 1)^{\alpha}$ × $\max(S_{cr}/\kappa, 1)^{-0.601}$ × $\min(S_{cys}/0.8, 1)^{-0.375}$ × $\max(S_{cys}/0.8, 1)^{-0.711}$ × 0.995^{Age} × 0.969 [if female] × 1.08 [if black] κ = 0.7 (females) or 0.9 (males) α = −0.248 (females) or −0.207 (males) $\min(S_{Cr}/\kappa$ or 1) indicates the minimum of S_{cr}/κ or 1 $\max(S_{cr}/\kappa$ or 1) indicates the maximum of S_{cr}/κ or 1 $\min(S_{cys}/0.8, 1)$ indicates the minimum of $S_{cys}/0.8$ or 1 $\max(S_{cys}/0.8, 1)$ indicates the maximum of $S_{cys}/0.8$ or 1
6) Bedside Schwartz
eGFR = 0.413 × (height/S_{cr}) if height is expressed in centimeters OR 41.3 × (height/S_{cr}) if height is expressed in meters
7) CKiD
eGFR = 39.8 × $[ht/S_{cr}]^{0.456}$ × $[1.8/Scys]^{0.418}$ × $[30/BUN]^{0.079}$ × $[1.076^{male}]$ $[1.00^{female}]$ × $[ht/1.4]^{0.179}$ where BUN is blood urea nitrogen in mg/dL Scys is cystatin C in mg/dL ht (height) is in meters S_{cr} is standardized serum creatinine in mg/dL

CKD, chronic kidney disease; *CKD-EPI*, chronic kidney disease epidemiology collaboration; *CKiD*, chronic kidney disease in children; *eGFR*; estimate glomerular filtration rate; *GFR*, glomerular filtration rate; *MDRD*, modification of diet in renal disease.

Table 35.4. The equations and their properties are discussed below.

Creatinine-based estimating equations

Cockcroft–Gault equation

The Cockcroft–Gault (C–G) equation was published in 1976 and estimates creatinine clearance in milliliter per minute from serum or plasma creatinine, age, and body weight. There is an adjustment factor for women that is based on a theoretical assumption of 15% lower creatinine generation because of lower muscle mass. Comparison to normal values requires calculation of the patient's BSA and adjustment to 1.73 m^2.

The C–G equation has several key limitations. First, it estimates creatinine clearance rather than GFR; hence, it is expected to overestimate GFR. Second, inclusion of a term for weight in the numerator causes the formula to systematically overestimate creatinine clearance in patients who are edematous or obese and will produce changes in estimated creatinine clearance with weight loss. Third, it was derived using older assay methods for serum creatinine, which had different calibration than newer assay methods that causes

positive bias with creatinine results from methods with calibration traceable to IDMS. Fourth, its age term leads to a sharp and excessive decline in estimated creatinine clearance with age. As a result, it has been shown to be substantially less accurate than modification of diet in renal disease (MDRD) study or CKD-EPI equations (see below).

Modification of diet in renal disease study equation

The MDRD study equation uses serum creatinine, age, sex, and race (black versus white or other) to estimate GFR as measured by urinary clearance of ^{125}I-iothalamate. The original six-variable equation also included serum urea and albumin; however, the six-variable equation is not used in clinical practice, because it is not expressed for creatinine methods with calibration traceable to IDMS. The four-variable equation has been reexpressed for use with serum creatinine methods with calibration traceable to IDMS [21]. The four-variable equation has been validated in African Americans, people with diabetic kidney disease, and kidney transplant recipients, three groups not included in large numbers in the original MDRD study. The equation underestimates GFR in populations with GFR of >60 mL/min/1.73 m^2, such as patients with type 1 diabetes without albuminuria and people undergoing kidney transplant donor evaluation. Because of limitations in accuracy of the MDRD study equation at higher values, GFR estimates are reported as a numeric value only if eGFR is <60 mL/min/1.73 m^2 and as ">60 mL/min/1.73 m^2" for higher values. The MDRD study equation has not been validated in children, pregnant women, or the very old (age $>$ 85 years). The MDRD study equation has demonstrated less variability and greater overall accuracy than the C−G formula [12].

Chronic kidney disease epidemiology collaboration equation

The CKD-EPI 2009 creatinine equation, was developed from a large database of individuals from research studies and patients from clinical populations with diverse characteristics, including people with and without kidney disease, diabetes, and a history of organ transplantation [22]. The equation is based on the same four variables as the MDRD study equation, but uses a two-slope "spline" to model the relationship between GFR and serum creatinine that improves the underestimation of GFR at higher values observed with the MDRD equation. As a result, the CKD-EPI equation is as accurate as the MDRD equation at eGFR <60 mL/min/1.73 m^2 and more accurate at higher values. The CKD-EPI equation has somewhat better precision (less scatter in values), although either estimating equation has substantial scatter in the agreement of individual eGFRs vs. measured GFR due to limitations discussed in the next section. The CKD-EPI equation was validated for a wide range of characteristics, including age, sex, race, body mass index, and presence or absence of diabetes or history of organ transplantation. A review of 12 assessments of the CKD-EPI and MDRD equations indicated that the CKD-EPI equation gave more accurate results at eGFR >60 mL/min/1.73 m^2 and the MDRD gave better performance at lower eGFR. The authors concluded the CKD-EPI equation was preferred for a general practice and public health perspective [23]. Differences between the CKD-EPI vs. MDRD study equations lead to marked differences in the distribution of eGFR values in the U.S. National Health and Nutrition Examination Survey data, with a lower prevalence of eGFR <60 ml/min/1.73 m^2. Similar results have been observed in multiple studies around the world. The differences between equations also lead to differences in multivariable associations with prognosis for adverse events as shown in a meta-analysis of 25 general population cohorts of 940,366 people [24]. Reclassification among eGFR categories also gave more accurate risk prediction. That is, patients reclassified to lower categories (higher GFR) using the CKD-EPI equation had a lower adjusted risk for mortality, CVD mortality and kidney failure than those not reclassified. Conversely, subjects reclassified to a lower eGFR category had a higher risk for the aforementioned disease states. With the CKD-EPI equation, it is now possible to report eGFR for adults across the entire range of values with minimal bias and reasonable imprecision. The U.S. National Kidney Foundation and the KDIGO 2012 CKD guideline recommend using the CKD-EPI equation and most large reference laboratories in the United States have transitioned to use the CKD-EPI equation (in place of the MDRD study equation). Reporting eGFR values above 60 mL/min/1.73 m^2 allows clinicians to identify patients who are at high risk for CKD before they develop reduced GFR, and to follow the progression in those with known CKD (i.e., polycystic kidney disease or patients with albuminuria).

Limitations of creatinine-based estimates of glomerular filtration rate

The variables included in any equation account for the average difference in creatinine generation from muscle mass due to the influence of age, sex, race, or body size but cannot account for other sources of variation in creatinine concentration that may be caused by other factors in individual patients. The estimate assumes an individual has "average" characteristics regarding body size and muscle mass. There will be more uncertainty in the eGFR to the extent an individual differs from "average." Table 35.5 lists patients for whom GFR estimates are likely to have large errors and clinical circumstances in which decisions based on inaccurate estimates may have adverse consequences.

TABLE 35.5 Indications for measured glomerular filtration rate when glomerular filtration rate estimates based on serum creatinine are likely to have large errors or may have adverse consequences.

- Extremes of age and body size
- Severe malnutrition or obesity
- Disease of skeletal muscle
- Paraplegia or quadriplegia
- Evaluation for kidney donation
- Vegetarian diet
- Before administration of prolonged courses of toxic medications

For example, no creatinine-based equation is expected to be accurate for patients with extremes of creatinine generation such as amputees, muscle builders, patients with muscle-wasting conditions, chronically ill or hospitalized patients who frequently have reduced muscle mass, or people with high or low dietary meat intake. In addition, current estimates are not sufficiently accurate to determine normal kidney function in a person who is being evaluated for a kidney donation. Finally, although eGFR is sufficient for adjustment of many medications, measured GFR should be considered before initiation of prolonged and potentially toxic therapy, such as cancer chemotherapy. When these situations occur, measured GFR or measured creatinine clearance may be helpful in clinical decision-making [25].

Cystatin-C-based estimating equations

The CKD-EPI 2012 cystatin C equation and the CKD-EPI 2012 creatinine and cystatin C combined equation were developed from a diverse set of studies, similar to that used for development of the CKD-EPI creatinine equation [22]. The cystatin C equation had similar accuracy to the creatinine equation, whereas the equation that used both creatinine and cystatin C was more accurate than an equation that used either cystatin C or creatinine alone, with the improvement due to better precision rather than bias. There is similar higher accuracy of the combined creatinine and cystatin C equation in subgroups with a trend towards higher accuracy of eGFR from cystatin C vs. estimated GFR from creatinine for subgroups with lower BMI. Other studies have confirmed the findings that the combined equation is more precise than eGFR from creatinine or cystatin C. Based on a small number of studies, performance of the CKD-EPI 2012 cystatin C equations are better than the CKD-EPI 2009 creatinine equation in regions outside North America, Europe and Australia, requiring less or no modification by a local coefficient.

Despite the similarities of accuracy of estimating equations between creatinine and cystatin C, eGFR from cystatin C provides more precise risk associations than eGFR from creatinine, with the risk associations of eGFR from creatinine and cystatin C intermediate between the two. Reclassification among eGFR categories also showed large differences with the different equations. Overall, net reclassification index favored eGFR from creatinine and cystatin C over eGFR from creatinine for all cause and cardiovascular mortality, but no difference was observed for ESRD [26].

Estimating glomerular filtration rate in children

Children have lower concentrations of creatinine than adults. Consequently, the influence of bias, imprecision, and interfering substances on creatinine results has a larger proportional effect. Younger children frequently have lower total protein than adults, with potentially larger negative biases caused by over-correction (over-compensation) for "average" protein concentrations in some Jaffe methods. Very young children may have elevated bilirubin and hemoglobin F, which can interfere in many Jaffe and enzymatic procedures. The relative imprecision [coefficient of variation (CV) %] of all creatinine methods becomes larger at lower concentrations. Creatinine methods with lower imprecision and fewer interferences should be used for pediatric populations. Equations to estimate GFR in children have used serum creatinine and height, which is a good approximation of muscle mass and BSA in children. The bedside Schwartz equation is recommended to estimate GFR if only creatinine is available [27].

Cystatin C is a promising biomarker for use in children because it is not influenced by muscle mass. The CKiD study has developed an equation using cystatin C, BUN and creatinine [28]. However the cystatin C used in development of the equation had not been standardized and therefore the accuracy of the estimating equation is not known when used in different clinical laboratories. An IFCC working group has reported a cystatin C based GFR estimating equation that used a standardized cystatin C method and was validated with a cohort of 262 Dutch and 440 Swedish children [29]. The equation has not been validated for use with a broader demographic mix.

Kidney function assessment and drug dose adjustment

Before standardized calibration of creatinine methods, there was substantial variability in methods used to assay serum creatinine in different clinical laboratories and in pharmacokinetic (PK) studies to establish drug dose adjustment on the basis of kidney function. The results from the PK studies were incorporated into U.S. Food and

Drug Administration (FDA) drug labels. The recommended drug dosages based on PK studies (i.e., the FDA drug labels) have been inconsistently translated into clinical practice. Since the introduction of standardized creatinine methods approximately 10 years ago, there is now less variation in estimating kidney function and more consistent drug dosing.

Laboratories and clinical practitioners are now reporting eGFR determined by the MDRD or CKD-EPI equations using standardized creatinine methods that give the most accurate assessment of kidney function, whereas the older drug labeling recommendations for dosing based on C−Gderived estimated creatinine clearance (or creatinine concentration in some cases) generally using nonstandardized methods. A large simulation study compared drug-dosing recommendations based on eGFR using the MDRD equation and estimated creatinine clearance using the C−G equation, both calculated from standardized creatinine values, to dosing recommendations based on gold-standard measurements of GFR [30]. The results suggested that for most patients and for most drugs tested, there was little difference in the drug dose that would be administered using either equation to estimate kidney function.

On the basis of these and other considerations, a KDIGO controversies conference recommended to use the most accurate method available to assess kidney function when evaluating the GFR for drug dosing [31]. Because standardization of creatinine measurement procedures has caused creatinine values to decrease vs. older methods, the most accurate method will not be the C−G equation. Estimated GFR from CKD-EPI equations should be used (creatinine or creatinine-cystatin C). Because the CKD-EPI (or MDRD study) equations are expressed as adjusted for BSA, in very large or very small patients the reported eGFR should be multiplied by the estimated BSA to obtain eGFR in units of milliliters per minute for drug dose adjustment. The NKDEP also recommends to use a measured GFR or creatinine clearance for situations when estimates based on creatinine are likely to be incorrect (see Table 35.5) or for drugs with narrow therapeutic or toxic thresholds. Over the past 10 years, an increasing number of studies have used methods other than C−G for assessment of drug dosage.

Urine protein

Definitions

Plasma contains approximately 300 g of protein; yet, the final urine filtrate is virtually protein-free because of selectivity of glomerular filtration. This conservation of essential proteins is necessary for oncotic regulation, for immune protection, for normal coagulation, and other vital processes. Proteinuria refers to increased excretion of any urinary protein, including albumin and other serum proteins, and proteins synthesized by the tubule (e.g., Tamm−Horsfall glycoprotein) or in the lower urinary tract. Albuminuria refers to increased excretion of albumin. An increased protein excretion rate is usually due to kidney damage. The specific proteins excreted depend on the type of kidney disease. Increased excretion of albumin is a sensitive and specific marker for CKD due to diabetes, hypertension, and glomerular diseases. Therefore tests for albuminuria are preferred for detection of CKD because it is more sensitive and specific for the common causes of CKD.

Mild proteinuria does not generally cause clinical signs or symptoms, but when severe (>3 g/day), consequences such as hypoalbuminemia, hypercholesterolemia, hypercoagulability, and hemodynamic instability may occur. This constellation of signs and symptoms is referred to as nephrotic syndrome. Recognition of nephrotic-range proteinuria or nephrotic syndrome may help identify the etiology of the kidney disease. In addition, the degree of proteinuria is a risk factor for kidney disease progression. However, even the presence of mild elevations in albuminuria (e.g., $<$30mg/g creatinine) is an independent risk marker for CVD, particularly in patients with diabetes, hypertension, or advanced age.

The threshold for abnormally elevated concentrations depends on what is being measured (total protein or albumin) and the collection method. A 24-h collection has been regarded as the gold standard but is difficult to implement in routine practice and not generally used. Measurement of albumin or total protein concentration in a randomly collected sample avoids the need for collection of a timed urine specimen but is affected by the state of hydration. Taking the ratio of albumin- or total protein-to-urine creatinine concentrations minimizes hydration as a source of variation but is affected by urine creatinine excretion rate. Table 35.6 shows the thresholds for the definition of albuminuria and proteinuria according to collection method and sex where appropriate. Measurement of the ratio of albumin-to-creatinine or total protein-to-creatinine in a random (also called "spot") urine sample is recommended [32]. A first morning urine void has been shown to have comparable diagnostic utility to a 24-h collection when expressed as the ACR [33]. Both total protein and albumin excretion can be increased transiently because of several factors, including urinary tract infection, hemodynamic stress (exercise, fever, and congestive heart failure), and transient metabolic perturbations (ketosis and hyperglycemia), which emphasizes the importance of testing only in high-risk individuals who have a high pretest probability of disease and for repeated testing to confirm the diagnosis of CKD. Thus, to confirm persistence of proteinuria, two or three abnormal tests are required over a minimum of 3 months [2]. Repeat confirmation testing should be performed from a first morning

TABLE 35.6 Thresholds for the definition of albuminuria and proteinuria according to collection method and sex.

Measure	Categories		
	Normal to mildly increased	Moderately increased	Severely increased
AER (mg/24 h)	<30	30–300	>300
PER (mg/24 h)	<150	150–500	>500
ACR			
(mg/mmol)	<3	3–30	>30
(mg/g)	<30	30–300	>300
PCR			
(mg/mmol)	<15	15–50	>50
(mg/g)	<150	150–500	>500
Protein reagent strip	negative to trace	negative to +	+ or greater

ACR, Albumin-to-creatinine ratio; *AER*, albumin excretion rate; *PER*, protein excretion rate; *PCR*, protein-to-creatinine ratio.
Source: Reprinted with permission from Chapter 94 in the Oxford Textbook of Clinical Nephrology (fourth ed.), N. Turner, N. Lameire, D.J. Goldsmith, C.G. Winearls, J. Himmelfarb, G. Remuzzi (Eds.), Oxford University Press, 2015.

collection to reduce the frequency of falsely elevated urine albumin [34].

Urine total protein measurement

Urine total protein is most frequently measured as turbidity after denaturation with trichloroacetic acid, sulfosalicylic acid, or benzethonium chloride; or a dye-binding reaction with pyrogallol red, molybdate or pyrocatechol-violet molybdate. Reactions with benzethonium chloride or the dye-binding reagents are the most common automated procedures. Each of these reactions gives a different response to different proteins, with albumin generally having the largest response. In addition, urine contains a large number of different proteins and other matrix solutes that contribute to variable responses among methods for total protein. There is no JCTLM-listed reference measurement procedure or reference material for urine protein. Because of these variables, there are large differences in measured results among methods for total protein that are not likely to be improved. Consequently, using generalized decision criteria to interpret urine protein will give different disease classifications based on the method used.

For patients with advanced kidney damage, large amounts of protein (>1 g/L albumin and others) are present in urine and will be detected by any method. Urine protein "dipsticks" typically detect approximately 300 mg/L of albumin and may not detect similar concentrations of other proteins such as Tamm–Horsfall glycoprotein associated with tubular damage. Urine protein is relatively insensitive to detect early kidney damage, and specific methods for urine albumin should be used for screening at-risk populations and for monitoring progression of kidney disease.

Urine albumin measurement

Urine albumin is the preferred analyte to detect early kidney damage and to monitor progression of CKD. Increased concentrations can be caused by excess glomerular leakage or by reduced tubular uptake. Albumin exists in plasma and urine as a heterogeneous molecule [32]. Plasma albumin is glycated, phosphorylated, nitrosylated, oxidized, binds many different ligands, and undergoes proteolytic modifications and truncation of C- and N-terminal amino acids. In addition, several genetic variants have been reported. Proteolysis also occurs in the tubules, bladder, and urine after collection. Urine concentrates many ligands, which alters the binding ratios, and urine itself has a widely varying matrix of solutes that can affect an antibody's binding reaction. Large (>5 kD) and small (500–5000 Da) fragments of albumin exist in plasma and in urine. Tubular uptake is receptor-mediated and may influence enrichment of modified plasma forms of albumin in urine. For example, glycated forms are a greater proportion in urine because of decreased uptake in the tubules.

Urine albumin is almost exclusively measured by turbidimetric or nephelometric immunoassay with typical limits of quantification being approximately 5 mg/L. Polyclonal and monoclonal antibodies are used, and different antibodies target different epitopes. The influence of molecular

forms of albumin on some immunoassays' responses has been investigated for a few partially cleaved and other conditions. Polyclonal antibodies were generally more efficient than monoclonal antibodies to recognize some modified albumin forms [32]. Urine albumin measured in 340 frozen individual urine samples using an LC-IDMS procedure found the same concentrations when quantitated from trypsin fragments near the amino- or carboxy-terminus or middle peptide regions of the molecule [35]. Measurement of urine albumin in 332 individual nonfrozen urine samples using 16 commercially available urine albumin immunoassay procedures found that bias was the dominating error component with differences in median values among methods of approximately 40% and CV from analytical error components of <5.8% for 14 of 16 methods [36]. In the same study, specimen-specific influences were 3.1%−6.6% for 12 of 16 methods, 13%−14% for three platforms from the same manufacturer, and could not be determined for one manufacturer due to analytical imprecision CV of 12.7%. These reports suggest that there was minimal influence of urine matrix or albumin molecular forms for most immunoassay methods. In addition, the results demonstrate that improvement in calibration standardization is needed for routine urine albumin methods to better utilize the clinical guidelines decision values for ACR.

Several reports used size-exclusion high-performance liquid chromatography (HPLC) to measure urine albumin and observed higher values than for immunoassays. These observations lead to a hypothesis that there were "nonimmunoreactive" forms of albumin in urine of diabetics. This hypothesis is no longer accepted on the basis of results such as the area under receiver-operator characteristic (ROC) curves for immunoassay being superior to HPLC for predicting cardiovascular events in diabetics [37] and demonstration that the increased values by HPLC were caused by coeluting proteins being erroneously quantified as albumin [32].

There is no JCTLM-listed reference measurement procedure or reference material for urine albumin. Most methods trace calibration to ERM-DA470, or its replacement ERM-DA470k/IFCC, Human Serum Proteins (JRC, EU) certified reference material. This reference material is intended for serum measurements and is diluted to achieve albumin concentrations measured in urine. Dilution protocols and diluent matrix are usually not specified, and it is not known if the diluted reference material is commutable with patient urine samples. Work is in progress to develop an IDMS-based reference measurement procedure and reference materials suitable for use for urine albumin measurements. NIST has released SRM 2925 pure human albumin intended for use to calibrate IDMS reference measurement procedures but not intended to calibrate immunoassays. NIST is preparing SRM 3666 albumin in frozen human urine that will have four concentrations of albumin and is intended for use to calibrate immunoassays. NIST, the Mayo Clinic, and the University of Minnesota are collaborating to develop LC-IDMS reference measurement procedures for urine albumin [38].

Urine creatinine measurement

Urine creatinine is measured using the same reagent systems as for serum or plasma, and the same reference measurement procedures have been used for urine samples. The urine matrix is lower in protein than serum and may have more concentrated solutes that may influence selectivity for creatinine. There is little information available on harmonization among results measured using different methods for urine creatinine, although the measurement is commonly made on urine samples as a correction for hydration and urine excretion rate influences on analyte concentrations. NIST has released SRM 3667 creatinine in frozen human urine for use to calibrate urine creatinine methods. SRM 3666 albumin in frozen human urine will have certified values for creatinine as well as for albumin.

Recommendations for measuring and reporting urine albumin and creatinine

An expert group of the NKDEP and IFCC recognized that standardization of urine albumin and creatinine measurement and reporting was needed, and that additional information was required to accomplish these goals [32]. The following recommendations were made regarding current practice for measuring and reporting urine albumin:

- A first morning urine sample has lower biologic variability than a random collection and is preferred. Twenty-four-hour urine collections are not needed and introduce collection errors.
- Albumin should be measured on fresh (nonfrozen) urine.
- Urine should be frozen at $\leq -70^{\circ}C$ (not $-20^{\circ}C$) when stored for investigational purposes.
- Urine ACR should always be reported. The reporting units should be albumin/creatinine as milligram per gram or milligram per millimole and be used uniformly in a country or region. The ACR should also be reported when a 24-h urine collection is used, because the interpretive guidelines are based on ACR, and complete 24-h collections are challenging.
- Albumin concentration (e.g., as milligram per liter) should not be reported by itself, because it is difficult to interpret without information on urine excretion rate.
- The term "urine albumin" should replace the term "microalbumin" as the name for the laboratory test, because the latter term is confusing to clinicians who are unsure of its meaning.

Experience showed that laboratories used inconsistent reporting practices for highly elevated values for urine albumin, ACR, total urine protein, and protein/creatinine ratio. Because physicians need quantitative values for highly elevated concentrations to appropriately monitor progression and response to therapy for advanced kidney disease, the NKDEP published updated recommendations for reporting these biomarkers [39]. The recommendations are to perform dilutions of urine containing elevated concentrations of albumin and/or total protein, and to report quantitative values. When the concentration of urine albumin or total protein exceeds the measuring capability after dilution, a greater than value for the concentration and for the ratio to creatinine should be reported based on the highest concentration that can be measured after dilution. It is not appropriate to simply report a value greater than the nominal analytical measuring range for undiluted urine samples.

Other biomarkers of kidney damage

Markers of tubular function

The tubules are active participants in homeostasis of minerals and other compounds. Therefore damage to the tubules can lead to abnormalities of potassium, bicarbonate, glucose, phosphate, amino acids, calcium, and water. It is more difficult to measure specific functions of the tubules than it is to measure GFR. Due to the adaptive nephron hypothesis, GFR decreases and so does the ability of the tubules to adapt to changes in the milieu. In addition, there are abnormalities that can be observed even at high levels of GFR with diseases that specifically affect the tubules. Abnormalities can be due to genetic abnormalities, changes in dietary intake of key nutrients, medications or toxins, infiltrative diseases in the kidney (e.g., cancer), or systemic diseases (e.g., rheumatic diseases). Thus abnormalities in any of these compounds might indicate tubular damage.

Urinalysis

Multiparameter dipstick urinalysis and microscopic examination are covered in Chapter 37, Water and electrolyte balance. Noting the presence of cellular components in urine is an indication of kidney damage. Cellular components present in urine in two or more examinations over at least 3 months can be diagnostic for CKD. Identification of the presence or absence of hematuria and proteinuria, as well as casts, may facilitate identification of an etiology for kidney disease. In particular, presence of hematuria, proteinuria, and cellular casts indicate glomerular disease, which often requires specific treatments.

Acute kidney injury and emerging biomarkers

Acute kidney injury (AKI) is defined by an increase in serum creatinine of 0.3 mg/dL or greater within 48 h, an increase in serum creatinine of 1.5-fold or greater above baseline that is known or presumed to have occurred within 7 days, or urine volume of <0.5 mL/kg/h for 6 h or more. Acute kidney disease is defined as kidney disease that is thought to have been present less than 90 days. AKI may occur in various settings. Most commonly, it occurs after exposure to hemodynamic insults (e.g., with massive volume depletion or after surgery) or after administration of nephrotoxic pharmaceutical agents (e.g., radiographic contrast, chemotherapy, or antimicrobials).

It is difficult to estimate an acute loss of kidney function from serum (or plasma) concentrations of the endogenous filtration markers creatinine or cystatin C, or from eGFR derived from equations, because the changes in serum concentrations lag behind the change in measured GFR. For example, after an abrupt decline in GFR, the serum concentration of creatinine will return to steady state after 2−3 days. In conditions in which the GFR is constantly changing (higher or lower), it will be even more difficult to estimate the GFR from serum concentrations of creatinine or cystatin C. The eGFR is important in managing patients with AKI because of the need for appropriate dosing of medications, and therefore estimation of GFR requires an analysis of the rate of change in eGFR and the absolute value at baseline.

To overcome the limitations of the serum filtration markers, there has been interest in novel biomarkers for early detection of kidney damage prior to the decrease in kidney function. AKI biomarkers provide evidence for tubule function, distinguish etiologies of AKI (particularly prerenal disease from acute tubular necrosis), or identify subclinical AKI that does not lead to decreased GFR but does portend a poor prognosis. Most AKI biomarkers are not ready for widespread clinical practice at this time. Only one AKI biomarker has been FDA cleared for clinical use. The NephroCheck test measures urine insulin-like growth factor-binding protein 7 (IGFBP7) and tissue inhibitor of metalloproteinases-2 (TIMP-2). Both of these molecules are inducers of G_1 cell cycle arrest, a key mechanism implicated in AKI. The test multiplies the two results to obtain a value that provides information on risk of development of AKI in the next 12 h. In the original validation trial, the combined biomarkers demonstrated an area under the curve in ROC analysis of 0.80 (0.76 and 0.79 alone). Furthermore, the combination of the two ([TIMP-2] · [IGFBP7] $(ng/ml)^2/1000$) significantly improved risk stratification when added to a nine-variable clinical model that included age, serum creatinine, apache III score, history of hypertension, nephrotoxic drugs, liver disease, sepsis, diabetes, and history of CKD [40].

Neutrophil gelatinase-associated lipocalin (NGAL) is approved for clinical use in several countries but not in the United States. NGAL is a protein (25 kD) found in human neutrophils and has a role in kidney embryology. Distal tubule injury upregulates the gene for synthesis, and the protein is secreted into blood and urine where it can be measured. NGAL is reported to increase in 2–6 h after AKI. NGAL can be elevated following injury to various tissues other than kidney, which has made diagnostic thresholds for AKI challenging to establish and limited its general use.

Kidney stones

Renal calculi, commonly called kidney stones, are solid aggregates that form in the renal pelvis, ureter, or bladder. Kidney stones occur when excess amounts of poorly soluble substances, such as calcium, oxalate, phosphate, urate, cysteine, and others, due to nonkidney pathophysiology are filtered through the glomerulus and form precipitates or aggregates, as the urine is concentrated by water reabsorption. Urine contains a large number of minerals and organic molecules being eliminated from the body many of which approach saturation concentrations and can be precipitated by moderately abnormal concentrations or pH in the urine. The frequently observed precipitates in refrigerated urine specimens demonstrate this phenomenon. The most common stones are calcium oxalate or a mixture of calcium oxalate and phosphate. However, a large range of minerals, organic molecules, and proteins can be found in kidney stones.

Although kidney stones are not useful as kidney function biomarkers, an analysis of the urine and the stone composition is useful to identify what substances are present in elevated concentrations. Stone composition analysis is usually performed by specialty laboratories using techniques such as infrared spectroscopy and X-ray diffraction analysis. Serum tests for PTH, vitamin D, and other molecules involved in mineral metabolism are also useful to understand the underlying pathophysiology. Once identified, treatment is based on controlling the concentrations of the components of the stone by pharmacological or diet strategies and staying well hydrated to prevent excessive concentration of urine that is conducive to formation of the insoluble stones.

References

[1] H.G. Rennke, B.M. Denker, Renal Pathophysiology: The Essentials, third ed., Lippincott Williams & Wilkins, Philadelphia, 2009, p. 378.

[2] Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group, KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease, Kidney Int. 3 (Suppl.) (2013) 1–150. Available from: https://kdigo.org/guidelines/ckd-evaluation-and-management//. accessed 06.03.2020.

[3] National Kidney Foundation: K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. Am. J. Kidney Dis. 2002;39:S1–266.

[4] United States Renal Data System, 2015 USRDS annual data report: epidemiology of kidney disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases: Bethesda, MD, 2015.

[5] K. Matsushita, M. van der Velde, B.C. Astor, M. Woodward, A.S. Levey, P.E. de Jong, et al., Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis, Lancet 375 (2010) 2073–2081.

[6] G.L. Myers, W.G. Miller, J. Coresh, J. Fleming, N. Greenberg, T. Greene, et al., Recommendations for improving serum creatinine measurement: a report from the Laboratory Working Group of the National Kidney Disease Education Program, Clin. Chem. 52 (2006) 5–18.

[7] W.G. Miller, G.R.D. Jones, Estimated glomerular filtration rate; laboratory implementation and current global status, Adv. Chronic Kidney Dis. 25 (2018) 7–13.

[8] <https://www.kidney.org/news/national-kidney-foundation-american-society-clinical-pathology-leading-laboratories-and> (accessed 06.03.2020).

[9] L.A. Inker, B.C. Astor, C.H. Fox, T. Isakovva, J.P. Lash, C.A. Peralta, et al., KDOQI US commentary on the 2012 KDIGO clinical practice guideline for the evaluation and management of CKD, Am. J. Kidney Dis. 63 (2014) 713–735.

[10] L.A. Stevens, C. Huang, A.S. Levey, Measurement and estimation of kidney function, in: J. Himmelfarb, M.H. Sayegh (Eds.), Chronic Kidney Disease, Dialysis and Transplantation: Companion to Brenner & Rector's The Kidney, third ed., Saunders, Philadelphia, 2010, pp. 22–38.

[11] G.J. Schwartz, H. Wang, B. Erway, G. Nordin, J. Seegmiller, J.C. Lieske, et al., Multicenter laboratory comparison of iohexol measurement, J. Appl. Lab. Med. 2 (2018) 711–724.

[12] L.A. Stevens, J. Coresh, T. Greene, A.S. Levey, Assessing kidney function—measured and estimated glomerular filtration rate, N. Engl. J. Med. 354 (2006) 2473–2483.

[13] N. Greenberg, W.L. Roberts, L.M. Bachmann, E.C. Wright, R.N. Dalton, J.J. Zakowski, et al., Specificity characteristics of seven commercial creatinine measurement procedures using enzymatic and Jaffe method principles, Clin. Chem. 58 (2012) 391–401.

[14] A. Boutten, A.-S. Bargnoux, M.-C. Carlir, P. Delanaye, E. Rozet, V. Delatour, et al., Enzymatic but not compensated Jaffe methods reach the desirable specifications of NKDEP at normal levels of creatinine. Results of the French multicentric evaluation, Clin. Chim. Acta 419 (2013) 132–135.

[15] N. Jassam, C. Weykamp, A. Thomas, S. Secchiero, L. Sciacivelli, M. Plebani, et al., Post-standardization of routine creatinine assays: are they suitable for clinical applications, Ann. Clin. Biochem. 54 (2017) 386–394.

[16] E. Lee, C.P. Collier, C.A. White, Interlaboratory variability in plasma creatinine measurement and the relation with estimated glomerular filtration rate and chronic kidney disease diagnosis, Clin. J. Am. Soc. Nephrol. 12 (2017) 29–37.

[17] A.A. Killeen, E.R. Ashwood, C.B. Ventura, P. Styer, Recent trends in performance and current state of creatinine assays, Arch. Pathol. Lab. Med. 137 (2013) 496–502.

[18] A.S. Levey, L.A. Inker, J. Coresh, GFR estimation: from physiology to public health, Am. J. Kidney Dis. 63 (2014) 820–834.
[19] J.H. Eckfeldt, A.B. Karger, W.G. Miller, G.P. Rynders, L.A. Inker, Performance in measurement of serum cystatin C by laboratories participating in the College of American Pathologists 2014 CYS Survey, Arch. Pathol. Lab. Med. 139 (2015) 888–893.
[20] A.S. Bargnoux, L. Pieroni, J.P. Cristol, N. Kuster, P. Delanaye, M.C. Carlier, et al., Multicenter evaluation of cystatin C measurement after assay standardization, Clin. Chem. 63 (2017) 833–841.
[21] A.S. Levey, J. Coresh, T. Greene, L.A. Stevens, Y.L. Zhang, S. Hendriksen, et al., Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate, Ann. Intern. med. 145 (2006) 247–254.
[22] A.S. Levey, L.A. Stevens, C.H. Schmid, Y.L. Zhang, A.F. Castro, H.I. Feldman, et al., A new equation to estimate glomerular filtration rate, Ann. Intern. Med. 150 (2009) 604–612.
[23] A. Earley, D. Miskulin, E.J. Lamb, A.S. Levey, K. Uhlig, Estimating equations for glomerular filtration rate in the era of creatinine standardization: a systematic review, Ann. Int. Med. 156 (2012) 785–795.
[24] K. Matsushita, B.K. Mahmoodi, M. Woodward, J.R. Emberson, T. H. Jafar, S.H. Jee, et al., Chronic Kidney Disease Prognosis Consortium. Comparison of risk prediction using the CKD-EPI equation and the MDRD study equation for estimated glomerular filtration rate, JAMA 307 (2012) 1941–1951.
[25] L. Stevens, A. Levey, Measured GFR as a confirmatory test for estimated GFR: indications and interpretation, J. Am. Soc. Nephrol. 20 (2009) 2305–2313.
[26] M.G. Shlipak, K. Matsushita, J. Arnlov, L.A. Inker, R. Katz, K.R. Polkinghorne, et al., Cystatin C versus creatinine in determining risk based on kidney function, N. Engl. J. Med. 369 (2013) 932–943.
[27] G.J. Schwartz, D.F. Work, Measurement and estimation of GFR in children and adolescents, Clin. J. Am. Soc. Nephrol. 4 (2009). 1832-43.
[28] G.J. Schwartz, M.F. Schneider, P.S. Maier, M. Moxey-Mims, V. R. Dharnidharka, B.A. Warady, et al., Improved equations estimating GFR in children with chronic kidney disease using an immunonephelometric determination of cystatin C, Kidney Int. 82 (2012) 445–453.
[29] A. Grubb, M. Horio, L.O. Hansson, J. Björk, U. Nyman, M. Flodin, et al., Generation of a new cystatin C-based estimating equation for glomerular filtration rate by use of 7 assays standardized to the international calibrator, Clin. Chem. 60 (2014) 974–986.
[30] L.A. Stevens, T. Nolin, M. Richardson, et al., Comparison of Drug Dosing Recommendations Based on Measured GFR and Kidney Function Estimating Equations, Am J Kid Dis. 54 (1) (2009) 33–42.
[31] G.R. Matzke, G.R. Aronoff, A.J. Atkinson, W.M. Bennett, B.S. Decker, K.U. Eckardt, et al., Drug dosing consideration in patients with acute and chronic kidney disease—a clinical update from Kidney Disease: Improving Global Outcomes (KDIGO), Kidney Int. 80 (2011) 1122–1137.
[32] W.G. Miller, D.E. Bruns, G.L. Hortin, S. Sandberg, K.M. Aakre, M.J. McQueen, et al., Current issues in measurement and reporting of urinary albumin excretion, Clin. Chem. 55 (2009) 24–38.
[33] H.J. Lambers Heerspink, R.T. Gansevoort, B.M. Brenner, M.E. Cooper, H.H. Parving, S. Shahinfar, et al., Comparison of different measures of urinary protein excretion for prediction of renal events, J. Am. Soc. Nephrol. 21 (2010) 1355–1360.
[34] S.H. Saydah, M.E. Pavkov, C. Zhang, D.A. Lacher, M.S. Eberhardt, N.R. Burrows, et al., Albuminuria prevalence in first morning void compared with previous random urine from adults in the National Health and Nutrition Examination Survey, 2009-2010, Clin. Chem. 59 (2013) 675–683.
[35] J.C. Lieske, O. Bondar, W.G. Miller, L.M. Bachmann, A.S. Narva, Y. Itoh, et al., A reference system for urinary albumin: current status, Clin. Chem. Lab. Med. 51 (2013) 981–989.
[36] L.M. Bachmann, G. Nilsson, D.E. Bruns, M.J. McQueen, J.C. Lieske, J.J. Zakowski, et al., State of the art for measurement of urine albumin: comparison of routine measurement procedures to isotope dilution tandem mass spectrometry, Clin. Chem. 60 (2014) 471–480.
[37] M.J. McQueen, H.C. Gerstein, J. Pogue, J.F. Mann, S. Yusuf, Reevaluation by high-performance liquid chromatography: clinical significance of microalbuminuria in individuals at high risk of cardiovascular disease in the Heart Outcomes Prevention Evaluation (HOPE) study, Am. J. Kidney Dis. 48 (2006) 889–896.
[38] W.G. Miller, J.C. Seegmiller, J.C. Lieske, A.S. Narva, L.M. Bachmann, Standardization of urine albumin measurements: status and performance goals, J. Appl. Lab. Med. 2 (2017) 423–429.
[39] Miller W.G., Bachmann L.M., Fleming J.K., Delanghe J.R., Parsa A., Andrew S. et al. Recommendations for reporting low and high values for urine albumin and total protein. Clin Chem. 2018; tbd.
[40] E.A. Hoste, P.A. McCullough, K. Kashani, L.S. Chawla, M. Joannidis, A.D. Shaw, et al., Derivation and validation of cutoffs for clinical use of cell cycle arrest biomarkers, Nephrol. Dial. Transpl. 29 (2014) 2054–2061.

Self-assessment questions

1. In which of the following scenarios should the CKD-EPI creatinine equation not be used?
   a. pregnant women
   b. people with amputations
   c. people with CKD
   d. people with muscular dystrophy
      i. all of the above
      ii. i, ii, and iv
      iii. i, ii, and iii
      iv. i and ii
2. What is the effect of eating a vegetarian diet on the concentration of serum creatinine and measured GFR, respectively?
   a. lower and higher
   b. higher and higher
   c. lower and lower
   d. no change in either
3. What is the definition for CKD?
   a. GFR $<$ 60 mL/min/1.73 m^2 or evidence of kidney damage that is present for $\geq$ 3 months
   b. eGFR $<$ 60 mL/min/1.73 m^2 and proteinuria that is present for $\geq$ 3 months
   c. eGFR $<$ 60 mL/min/1.73 m^2 and evidence of kidney damage that is present for $\geq$ 3 months
   d. eGFR $<$ 60 mL/min/1.73 m^2 or evidence of kidney damage
4. The CKD-EPI equation results in improved performance for estimation of measured GFR compared with the MDRD study equation in which of the following groups?
   a. people with CKD
   b. amputees
   c. people with rapidly changing kidney function
   d. people with higher GFRs
5. Which of the following statements is not correct?
   a. eGFR based on the CKD-EPI equation had stronger associations of risk for ESRD, mortality, and CVD compared with the MDRD study equation.
   b. eGFR based on the CKD-EPI equation lead to lower prevalence of CKD than with the MDRD study equation.
   c. Patients reclassified to lower CKD categories by the CKD-EPI equation had a lower adjusted risk for all risk mortality, CVD mortality, and kidney failure than those not reclassified.
   d. There is no difference between the eGFR from the CKD-EPI and MDRD study equations for assessment of CKD prevalence and risk for adverse outcomes.
6. Which is the preferred method for testing for urine albumin?
   a. first morning specimen to test for urine ACR
   b. first morning specimen to test for urine albumin
   c. random spot urine to test for urine ACR
   d. timed urine specimen to test for urine ACR
7. True or false? Adverse outcomes of CKD include CVD, kidney failure, anemia, abnormalities of mineral metabolism, infections, frailty, and cognitive impairment.
   a. true
   b. false
8. NKDEP recommends to report eGFR because
   a. It is easier to interpret a patient's kidney function from the eGFR than from serum creatinine.
   b. Physicians do not know how to perform the calculation.
   c. Patients with CKD are inadequately identified by serum creatinine.
   d. It gives a correct assessment of kidney function in all pathologic conditions.
      i. i and ii
      ii. i, ii, and iv
      iii. i, ii, and iii
      iv. i and iii
9. Measurement of serum cystatin C
   a. is clearly superior to serum creatinine as a marker of kidney function
   b. may be a useful biomarker in conditions when serum creatinine is a less reliable indicator of kidney function
   c. is standardized so any method can be used with any equation to calculate eGFR
   d. should become part of the basic metabolic profile

Answers

1. b
2. a
3. a
4. d
5. d
6. a
7. a
8. d
9. b

Chapter 36

Contemporary practice in clinical chemistry: blood gas and critical care testing

John Toffaletti
Duke University Medical Center, Durham, NC, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Describe the physiological factors that regulate pH, pCO_2, pO_2, and the electrolytes.
- Interpret pH, pCO_2, and bicarbonate (total CO_2) results to evaluate ventilatory and metabolic acid–base disorders, and if appropriate compensation is present.
- Describe the binding of oxygen to hemoglobin and the factors that affect this binding.
- Evaluate pO_2 results to determine the adequacy of arterial oxygenation and calculate the alveolar–arterial pO_2 difference.
- Interpret cord blood data to identify possible hypoxia and/or acidosis in the neonate.
- Describe the clinical utility of blood lactate and glucose results in critical care.
- Describe the proper collection and handling of specimens for blood gas and critical care testing.

Introduction

The term "blood gases" refers to the parameters pH, pCO_2, and pO_2, which are commonly measured in blood. Note that the little "p" in pH stands for negative log, while the italicized *p* in pCO_2 and pO_2 stands for the partial pressure of each of these gases. Blood gases are commonly measured by electrochemical (potentiometric or amperometric) methods that use ion-selective or gas-selective electrodes. This chapter uses the unit mmHg for blood gases, because it is readily understood and because many journals use mmHg as the conventional and SI unit. To convert mmHg into kPa, multiply the value by 0.1333. Cooximetry refers to the measurement of various forms of hemoglobin (Hb), including oxyhemoglobin (O_2Hb) and carboxyhemoglobin (COHb), from which the term is derived. Modern cooximeters are specialized spectrophotometers that measure the absorbance of over 100 wavelengths over the range of 478–672 nm. The absorbance spectrum is then analyzed to calculate the concentrations of various Hb forms present in blood: O_2Hb, COHb, metHb, and total Hb.

Explanations of blood gas, acid–base, and cooximetry terms

pH

pH is an index of the acidity or alkalinity of the blood. If normal arterial pH is 7.35–7.45 (45–35 nmol/L), then a pH $<$7.35 indicates an acidic state and a pH $>$7.45 indicates an alkaline state. In critical care, a clinically acceptable range of 7.30–7.50 is sometimes used as a guideline (see Table 36.1). Acidemia refers to the condition of blood being too acidic, and acidosis refers to the metabolic process within the patient that causes the acidemia; the adjective for the process is acidotic. Similar terms are used for the alkaline state: alkalemia, alkalosis, and alkalotic. All enzymes and most physiological processes are affected by pH.

pCO_2

pCO_2 is a measure of the tension or pressure of carbon dioxide dissolved in the blood. The pCO_2 of blood represents the balance between cellular production of CO_2 and the ventilatory removal of CO_2. A normal, steady pCO_2 indicates that the lungs are removing CO_2 from the blood at about the same rate as tissues are releasing CO_2 into the blood. A change in pCO_2 indicates an alteration in

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00036-3

TABLE 36.1 Reference ranges and critical limits for venous and arterial blood.

Measurement	Reference range (mixed venous)	Reference range (arterial)	Critical limits (arterial)
pH	7.33–7.43	7.35–7.45	≤7.25 or ≥7.60
pCO_2	41–51 mmHg (5.5–6.8 kPa)	35–45 mmHg (4.7–6.0 kPa)	≤20 or ≥60 mmHg (2.6 or 8.0)
HCO_3	21–30 mmol/L (21–30 mEq/L)	21–28 mmol/L (21–28 mEq/L)	≤10 or ≥40 mmol/L
pO_2	35–50 mmHg (4.7–5.3 kPa)	83–100 mmHg (11.1–13.4 kPa)[a]	≤ 40 mmHg (>10.7 kPa[a])
sO_2	70–75	96–100	< 85%
$\%O_2Hb$	–	94–99	< 85%
Anion gap	3–12 mmol/L	3–12 mmol/L	–
Base excess	−3 to +3	−3 to +3	–

pO_2, Oxygen pressure.
[a]*pO_2 of arterial blood varies with age.*

this balance, usually due to a change in the ventilatory status. CO_2 is an acidic gas, with blood levels largely controlled by our rate and depth of breathing or ventilation, provided that the rate of metabolic production of CO_2 is constant. pCO_2 contributes to the respiratory or ventilatory component of acid–base balance and is also a key factor in calculating the alveolar pO_2 (see the section "Evaluation of oxygenation status and arterial oxygenation").

pO_2

pO_2 is a measure of the tension or pressure of oxygen dissolved in the plasma of blood. Note that although the O_2 dissolved in plasma and the O_2 bound to Hb are in equilibrium, it is only the O_2 dissolved in plasma that is detected as pO_2 by the oxygen electrode. The pO_2 of arterial blood is primarily related to the ability of the lungs to oxygenate blood with alveolar air. A decreased arterial pO_2 indicates one or more of the following circumstances:

- Decreased pulmonary ventilation (hypoventilation), for example, as caused by airway obstruction or trauma to the brain.
- Impaired gas exchange between the alveolar air and the pulmonary capillary blood, for example, as caused by bronchitis, emphysema, pulmonary edema, or asthma.
- Altered blood flow within the heart or lungs, as caused by congenital defects in the heart or shunting of venous blood into the arterial system without oxygenation in the lungs.
- Intrapulmonary shunting when blood passes through nonfunctional or nonventilated alveoli.
- Pulmonary embolism in which blood flow to a segment of the lung is blocked by a blood clot.

Bicarbonate

Bicarbonate ion (HCO_3^-) may be measured directly by some blood gas analyzers or calculated [HCO_3^-] from the measurements of the pH and pCO_2. Although the terms CO_2, total CO_2, and bicarbonate are sometimes used interchangeably, the term "total CO_2" refers to the sum of the [HCO_3^-] and the dissolved CO_2 in plasma. Bicarbonate is an indicator of the buffering capacity of blood, with a low bicarbonate concentration indicating that a larger pH change will occur for a given amount of acid or base produced. Bicarbonate is classified as the metabolic component of acid–base balance.

Base excess

Base excess (BE) is a calculated term that describes the amount of bicarbonate relative to pCO_2. Some believe it helps to determine quickly the amount of bicarbonate that a patient may need and arguably may provide more useful information about acid–base status than the bicarbonate and pCO_2 [1]. It is calculated as follows:

$$BE = 0.929(HCO_3 - 24.4 + 14.8(pH - 7.4)) \quad (36.i)$$

It is calculated for hypothetical anemic blood (Hb = 5 g/dL) on the principle that blood Hb effectively buffers both the plasma and the much larger extracellular fluid. It is as if the blood Hb were dispersed in the larger pool of extracellular fluid. Because BE and the simple difference between 24 mmol/L (a normal bicarbonate) and the measured bicarbonate typically agree within 2 mmol/L, either provides virtually the same clinical information.

Anion gap

The anion gap (AG) is a calculated term for the difference between the commonly measured cations (Na and

TABLE 36.2 Changes of anion gap in various acid–base disorders.

Disorder	Decreased	Gained	Effect on AG
Diarrhea	HCO_3^-	Cl	Little
Renal tubular acidosis	HCO_3^-	Cl	Little
Lactate acidosis	HCO_3^-	Lactate	Increased
Ketoacidosis	HCO_3^-	Ketoacids	Increased
Mixed disorder: ketoacidosis with metabolic alkalosis	HCO_3^-	Ketoacids and HCO_3^-	Increased (with little change in total CO_2)

AG, Anion gap.

sometimes K) and the commonly measured anions (Cl^- and HCO_3^-). Therefore, the AG represents the unmeasured anions such as lactate, acetoacetate, and albumin. The AG is usually calculated as:

$$\mathrm{AG} = \left[\mathrm{Na}^+\right] - \left[\mathrm{Cl}^-\right] - \left[\mathrm{HCO_3}^-\right] \quad \left(\text{Reference Range: } 3-12 \text{ mmol/L}\right) \tag{36.ii}$$

If K is included in the calculation, the formula is:

$$\mathrm{AG} = \left[\mathrm{Na}^+\right] + \left[\mathrm{K}^+\right] - \left[\mathrm{Cl}^-\right] - \left[\mathrm{HCO_3}^-\right] \quad \left(\text{Reference Range: } 7-16 \text{ mmol/L}\right) \tag{36.iii}$$

Note: While K is a commonly measured cation and would be relevant to the AG calculation, it is usually not included in the calculation.

The AG is useful in diagnosing a metabolic acidosis and differentiating among the causes, as shown in Table 36.2. For example, in metabolic acidosis due to hypoxia, the blood lactate increases, which increases the AG. While blood lactate has become a common measurement, it is considered an "unmeasured" anion in the AG calculation. In addition, in patients with a low albumin, the AG will be lower by about 2.5 mmol/L for each g/dL fall in albumin [2,3]. Although often very useful, AG calculated from three measurements is subject to variations of up to ± 3 mmol/L and is also affected by the normal variation in albumin concentrations [2,3]. Reference ranges for the above parameters are given in Table 36.1.

Strong ion difference

The strong ion difference (SID) is a concept developed by Peter A Stewart in 1981 to explain changes in pH and acid–base balance, using electroneutrality, conservation of mass, and the degree of dissociation (strongly or weakly) of electrolytes [1]. It is complex mathematically, but in simplest terms, it is calculated as:

$$\mathrm{SID} = \left(\mathrm{Na}^+ + \mathrm{K}^+ + \mathrm{Ca}^{++} + \mathrm{Mg}^{++}\right) - \left(\mathrm{Cl}^- + \text{lactate}\right) \tag{36.iv}$$

A normal SID is approximately 40–45 mEq/L, with a decreased SID indicating an acidosis and an increased SID indicating an alkalosis. Because measurement of numerous analytes is required to calculate the SID, it is subject to a large analytical variation and the theory suggests that ionized concentrations are more appropriate. While SID may have value in selected acid–base disorders, there is no consensus on whether it improves the understanding or interpretation or treatment of acid–base disorders [4].

Hb and its derivatives

Hb and its derivatives are measured by cooximetry. These include O_2Hb, deoxyhemoglobin (HHb), COHb, and methemoglobin (metHb). Briefly, O_2Hb has four oxygen molecules bound to each of the heme groups in the Hb molecule. COHb has carbon monoxide (CO) bound very tightly to an O_2-binding site on a heme group, which also causes the remaining O_2 molecules to bind very tightly to the other three heme groups. COHb cannot release O_2 until the carbon monoxide is released. MetHb is inactive Hb (unable to bind O_2), because its ferrous ion (Fe^{2+}) has been oxidized to a ferric ion (Fe^{3+}). Hb function is discussed later in this chapter.

Percent O_2 saturation and $\%O_2Hb$

The percent O_2 saturation (sO_2) is the percentage of *functional* Hb (O_2Hb + HHb) that is saturated with oxygen. sO_2 is calculated as the concentration of O_2Hb divided by the functional Hb. Although sO_2 is clearly related to oxygenation in the lungs, pO_2 is somewhat better for assessing that function, because it increases linearly with higher O_2 levels, while the sO_2 levels off as it approaches 100%.

$$sO_2 = \frac{O_2Hb}{O_2Hb + HHb} \quad (36.v)$$

The $\%O_2Hb$ (formerly called the fractional Hb saturation) is the percentage of total Hb that is saturated with oxygen. The $\%O_2Hb$ may be used for determining the oxygen content and, therefore, the oxygen delivery (DO_2) and oxygen consumption (VO_2) [5].

$$\%O_2Hb = \frac{O_2Hb}{O_2Hb + HHb + COHb + MetHb} \quad (36.vi)$$

COHb and Met-Hb

COHb and metHb together normally make up only 1%–2% of the total Hb in blood. Neither is able to perform O_2 carrying and releasing functions. Exposure to carbon monoxide will increase the %COHb from about 1% to 5%–10% in smokers and to 50% or more with exposure to toxic or lethal levels. MetHb is Hb that has its Fe^{2+} ion oxidized to an Fe^{3+}, which also inactivates the ability of met-Hb to bind O_2.

DO_2 and VO_2

Both DO_2 and VO_2 are parameters for assessing critically ill patients. There are now several ways of determining these parameters, with laboratory measurements sometimes used. DO_2 in mL/min requires measurement of the following values in arterial blood: Hb in g/L, $\%O_2Hb$ from 0 to 1.0, pO_2 in mmHg (of minor importance in this calculation), and cardiac output (C.O.) in L/min.

$$\begin{aligned} DO_2 &= CO \times (1.34 \times Hb \times \%O_2Hb + 0.03pO_2) \\ &= C.O.(O_2a) \end{aligned} \quad (36.vii)$$

where O_2a is the arterial oxygen content.

VO_2 is oxygen consumption by the organs and tissues of the body. Put simply, it is the difference in oxygen content between the arterial (O_2a) and venous (O_2v) blood times C.O.:

$$VO_2 = C.O. \times (O_2a - O_2v) \quad (36.viii)$$

Physiology of acids and bases

How are acids and bases produced?

Metabolic acid

The vast majority of our metabolic acid is from CO_2 produced in the mitochondria as one of the ultimate byproducts of glucose oxidation. However, CO_2 is considered the "respiratory" component of our acid base balance, because lung ventilation directly affects the blood pCO_2 and can rapidly change the blood pH.

Lactate acidosis

Common causes of increased lactate are from ischemia or from mitochondrial damage due to factors (O_2 radicals, cytokines, etc.) produced by inflammation, as with sepsis. Contrary to common belief, lactic acid is virtually *never* produced. The only reaction in humans that produces lactate is when pyruvate is converted into lactate, a reaction that actually *consumes* acid by the participation of NADH being converted into NAD + [6]. During an oxygen deficit, the acid comes many biochemical steps later, when adenosine triphosphate (ATP) depletion causes accumulation of ADP, phosphate, and hydrogen ions. Normally, oxidative phosphorylation reconverts these products back to ATP with little net change of H ions. However, without oxygen, the ADP and acid accumulate to cause an acidosis.

Ketoacidosis

Diabetic ketoacidosis develops when a person has too little insulin such that glucose cannot enter cells to produce sufficient ATP energy. The liver begins to break down fat for energy, which produces toxic ketoacids formed by the deamination of amino acids and the breakdown of fatty acids. The two common ketones produced in humans are acetoacetate and β-hydroxybutyrate. Ketoacidosis can also occur with people undergoing starvation as the body is forced to break down fat for sustenance due to their lack of outside nutrition. In ketoacidosis, unregulated ketone production causes a severe accumulation of keto acids in the blood.

Production of base

Generally, there is little metabolic production of bicarbonate or other alkaline substances in the body. Bicarbonate is generated from metabolic CO_2 after it combines with H_2O and releases H^+. The kidney retains or excretes HCO_3^- depending on the physiologic need. Bicarbonate can increase if Cl is lost by either vomiting or by renal loss from diuretics and can decrease if excess Cl accumulates. Generally, any loss of acid will increase the proportion of HCO_3^- from CO_2. Bicarbonate given in excess to treat an acidosis is an iatrogenic source of base.

Buffer systems

Bicarbonate–carbon dioxide

Bicarbonate is the buffer in highest concentration (~24 mmol/L) in blood plasma, which is also of central importance in acid buffering in the blood. CO_2 is a volatile acidic gas, is soluble in water, and is the major acid produced as a product of energy metabolism. CO_2 produced by cell metabolism diffuses into blood (with a lower pCO_2), combines with H_2O to produce carbonic

acid (H_2CO_3), then immediately dissociates to bicarbonate and hydrogen ions. pH, HCO_3^- (in mmol/L), and pCO_2 (in mmHg) are related by the equation:

$$\mathrm{pH} = \mathrm{pK} + \log\left[\frac{\mathrm{HCO_3}^-}{0.03 \times p\mathrm{CO_2}}\right] \qquad (36.ix)$$

Note that the pK is defined as the pH at which the HCO_3^- and H_2CO_3 (or $0.03 \times pCO_2$) are in equal concentrations. That is, their ratio is 1, and the log of 1 is equal to 0.

At the normal concentration ratio in blood of 20:1 ("ideal" would be 1:1) and with a pK of 6.1 ("ideal" would be 7.4), the HCO_3^--H_2CO_3 buffer system would seem to be ill-suited to buffering pH in blood. However, this excess of base (HCO_3^-) along with the volatility of CO_2 comprises an efficient system for preventing accumulation of excess acid. The lungs effectively make this an open system, with the loss of CO_2 providing a large buffering capacity. It is the ratio of HCO_3^- to H_2CO_3 that determines the pH. Thus, HCO_3^-:H_2CO_3 in a concentration ratio of 15:0.75 has the same pH as a ratio of 20:1.0.

Hb

In addition to its oxygen binding and releasing properties, Hb also acts as a buffer by transporting acid from the tissues to the lungs. A remarkable feature of Hb is that it increases its affinity for hydrogen ions (H^+) as it loses oxygen. That is, HHb has a greater affinity for H^+ than does O_2Hb. In tissue capillaries, O_2Hb enters an environment of low pO_2 and high acid. These conditions promote release of O_2, which also promotes binding of H^+. In the lungs, HHb encounters an environment of high pO_2 and low acid, which promote gain of O_2 and release of H^+. These relationships are shown in Fig. 36.1.

Phosphate

The HPO_4^{2-}-$H_2PO_4^-$ buffer pair is of minor importance as a buffer in plasma, with a concentration of ~1 mmol/L (3.1 mg/dL). It is of greater importance, and in higher concentration, as an intracellular buffer.

Albumin and other proteins

Largely due to the imidazole groups on the amino acid histidine, with a pK of ~7.4, albumin and other proteins are also pH buffers. Because albumin is normally the major "unmeasured" anion in blood, albumin also affects the AG. For every 1 g/dL decrease in albumin below a normal level of ~4.4 g/dL, the AG will decrease by about 2.5 to 3 mmol/L. Patients in intensive care commonly have hypoalbuminemia, which can lower the AG and possibly confuse the interpretation of this parameter [2,8].

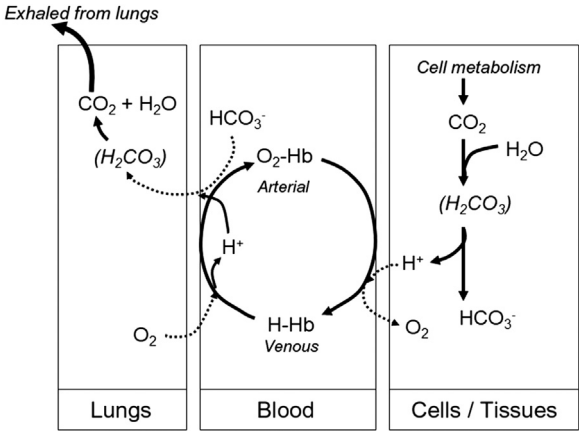


FIGURE 36.1 Interrelationships of the bicarbonate and hemoglobin buffering systems. Modified with permission from Ehrmeyer S.S. and Shrout J.B. Source*: From M.L. Bishop, J.L. Duben-Engelkirk, E.P. Fody (Eds.), Clinical Chemistry: Principles, Procedures, Correlations, second ed., Lippincott, Philadelphia, PA [7].*

Acid-base regulation

Respiratory (ventilatory) system

CO_2 is continually produced by aerobic metabolism, diffuses into blood, and then is removed by ventilation. Given that arterial CO_2 is influenced greatly by the ventilatory rate, pCO_2 is considered the respiratory component of the bicarbonate−CO_2 buffer system. Slight changes in H^+ ion concentration affect ventilatory rate, which is a major factor for pH regulation. Provided there is a sufficient gradient of pCO_2 between the tissues and the blood, CO_2 will readily diffuse into the blood, forming HCO_3^- and H^+ ions, with Hb readily accepting the H^+ (see Fig. 36.1). As Hb enters a region of high pO_2 in the lungs, HHb picks up O_2 to become O_2Hb, which immediately promotes loss of H^+. H^+ quickly combines with HCO_3^- to produce dissolved CO_2, which diffuses into the alveolar air for ventilatory removal.

The arterial pCO_2 represents a balance between tissue production and diffusion of CO_2 into blood, and pulmonary removal of CO_2. An elevated pCO_2 usually indicates inadequate ventilation (hypoventilation) and a respiratory acidosis. Conversely, a decreased pCO_2 usually indicates excessive ventilation (hyperventilation) and a respiratory alkalosis. There are several causes of respiratory abnormalities [9]. Respiratory acidosis (ventilatory failure) can be caused by the following:

- Obstructive lung disease (e.g., chronic bronchitis or emphysema)
- Impaired pulmonary blood flow (pulmonary embolism)
- Impaired function of respiratory center (head trauma, sedation, or anesthesia)
- Hypoventilation by mechanical ventilator

Respiratory alkalosis (hyperventilation) can be caused by:

- Hypoxemia (which stimulates hyperventilation)
- Anxiety
- Hyperventilation by mechanical ventilator
- Metabolic acidosis
- Septicemia
- Trauma

Metabolic (renal) system

The renal regulation of acid−base balance is complex, but in simple terms, when the H^+ concentration deviates from normal, the kidneys respond by reabsorbing or secreting H^+ and exchanging Cl^- with HCO_3^-, and other ion exchanges to regulate the pH of blood. Because kidneys are the major regulator of HCO_3^-, HCO_3^- is considered the metabolic component of the HCO_3^--CO_2 buffer system. (Although H^+ is regulated by the kidney, the liver, through gluconeogenesis, is more important than the kidney in removal of H^+.) Metabolic acidosis may develop either if H^+ accumulates or if HCO_3^- ions are lost. Metabolic alkalosis may develop from either loss of H^+ or increase of HCO_3^-. Although a change in ventilatory rate can alter arterial pH in minutes, the kidneys require hours to days to affect significantly pH by altering the excretion of HCO_3^-. Thus metabolic compensation takes several hours to alter the pH significantly.

Compensation

Compensation is a homeostatic response to an acid−base disorder in which the body attempts to restore pH to normal by adjusting the ratio of HCO_3^- to H_2CO_3 ($0.03 \times pCO_2$) back to a normal ratio of 20:1. Compensation involves either a relatively rapid ventilatory response (change in pCO_2) to a metabolic abnormality, or a relatively slow metabolic response (change in HCO_3^-) to a ventilatory abnormality. Some facts to remember about compensation are:

- It is driven by changes in pH. As compensation returns pH to normal, the pH-driven compensation process slows, then stops.
- Significant respiratory compensation of metabolic acidosis by loss of CO_2 occurs in 12−24 hours, while full metabolic compensation of respiratory acidosis by renal loss of HCO_3^- takes 2−5 days.
- Respiratory compensation by hyperventilation in metabolic acidosis is fairly predictable.
- While the respiratory response to metabolic alkalosis is hypoventilation, other factors such as pain and hypoxemia can stimulate hyperventilation and overcome the hypoventilatory effect of metabolic alkalosis [8].

The expected compensation for each acid−base abnormality will be discussed in the sections "clinical abnormalities of acid−base balance" and "Mixed acid−base disorders."

Clinical abnormalities of acid−base balance

Metabolic (nonrespiratory) acidosis

Metabolic acidosis is any clinical process that leads to a decreased blood pH (acidemia) and a decreased HCO_3^- level. It may be caused by gain of an acid or loss of a base, usually bicarbonate (increased AG), or a gain of Cl (normal AG). Causes of metabolic acidosis include the following [8,10]:

Acidosis is usually associated with little change in AG:

- Loss of HCO_3^- and gain of Cl: diarrhea and renal tubular acidosis (RTA). In RTA, insufficient HCO_3^- is reabsorbed in exchange for Cl, such that HCO_3^- is lost in the urine and Cl increases in the blood.

Acidosis is usually associated with an increased AG (Table 36.2):

- Excess production of acid from ketoacidosis, hypoxia, or lactate acidosis. Because insulin deficiency inhibits glucose entry into cells, to maintain adequate production of ATP, fatty acids are oxidized that produce ketoacids. Hypoxia results in more pyruvate conversion into lactate.
- Ingestion of acids or acid-producing substances: salicylates, ethanol, ethylene glycol, methanol, ammonium chloride, etc.
- Renal insufficiency with inadequate renal excretion of acids from normal metabolism.

Use of the anion gap in metabolic acidosis

The AG, especially when elevated, may be useful in diagnosing the type of metabolic acidosis and in indicating the possibility of a mixed acid−base disorder, as shown in Table 36.2.

Expected compensation in metabolic acidosis

In the initial phase of metabolic acidosis, both pH and HCO_3^- are decreased and pCO_2 is still normal (Table 36.3 and Fig. 36.2). Acidemia stimulates medullary chemoreceptors causing hyperventilation, and this respiratory compensation by hyperventilation should start to lower blood pCO_2 within minutes and have a significant effect by 2 hours, with maximal effect in 12 to 24 hours. Respiratory compensation typically lowers pCO_2 in a predictable manner (Table 36.3). If the pCO_2 is more than

TABLE 36.3 Expected compensation for acid–base disorders.

Metabolic Acidosis—CO_2 excretion by the lungs
pCO_2 decreases 1–1.3 mmHg for every 1 mmol/L decrease in HCO_3^-; pCO_2 is within 2 of sum $(1.5 \times HCO_3) + 8$ pCO_2 within 2 of 2 digits after decimal in pH
Metabolic alkalosis—CO_2 retention by the lungs
pCO_2 increases by 0.6–0.7 mmHg for every 1 mmol.L increase in HCO_3^-; pCO_2 is within 3 of sum $(0.65 \times HCO_3^-) + 23$
Respiratory acidosis—HCO_3 retention by kidneys
Acute: HCO_3^- increases 1 mmol/L for every 10 mmHg increase in pCO_2; Chronic: HCO_3^- increases 3–5 mmol/L for every 10 mmHg increase in pCO_2
Respiratory alkalosis—HCO_3 excretion by kidneys
Acute: HCO_3^- decreases by 1–2 mmol/L for every 10 mmHg decrease in pCO_2; Chronic: HCO_3^- decreases 3–5 mmol/L for every 10 mmHg decrease in pCO_2(minimum 12–14 mmol/L)

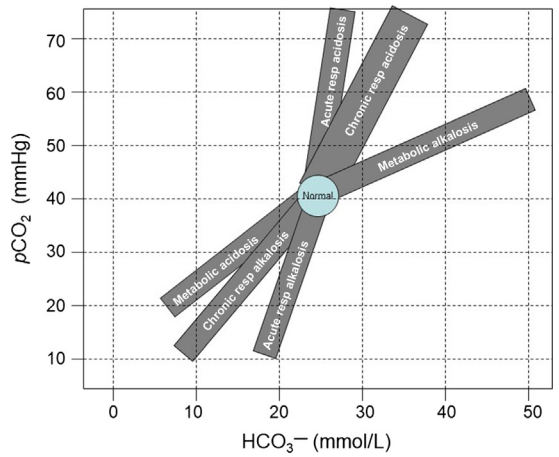


FIGURE 36.2 Expected relationships between HCO_3^- and pCO_2 during compensation for primary acid–base disorders. The areas shown are based on plots of equations shown in Table 36.3.

3 mmHg above the expected (calculated) value, the patient's lungs may not be fully capable of hyperventilating and may have an underlying respiratory acidosis. If the pCO_2 is less than 3 mmHg of this calculated value, an underlying respiratory alkalosis may be present.

Treatment of metabolic acidosis

The optimal treatment for metabolic acidosis is correcting the underlying cause, such as by administering insulin, improving DO_2, giving fluids as appropriate, correcting any electrolyte disorders, etc. If the condition must be corrected urgently, bicarbonate may be administered based on the blood bicarbonate level [8].

Metabolic (nonrespiratory) alkalosis

Metabolic alkalosis is an acid–base disorder characterized by an elevation in bicarbonate level above 26 mmol/L and an elevation in pH. Metabolic alkalosis is frequently associated with renal impairment, because a healthy kidney can excrete large amounts of HCO_3^- when it is in excess [1,10]. Causes of metabolic alkalosis include the following [1,8]:

- Loss of acid by vomiting or loss of acidic urine. The initial loss of gastric acid causes metabolic alkalosis, which may enhance renal tubular loss of K^+ to conserve H^+.
- Deficiency of Cl. Deficiency of Cl enhances HCO_3^- reabsorption in the renal tubule. Contraction of the intravascular volume can cause a Cl deficit and alkalosis ("contraction alkalosis").
- Hypokalemia. Hypokalemia stimulates distal tubular reabsorption of HCO_3^- ions.
- Administration of excess HCO_3^-. If excessive, HCO_3^- administration may lead to alkalosis, especially if renal function is compromised.
- Diuretics. Loop diuretics interfere with tubular reabsorption of Na and Cl, and thiazide diuretics enhance K loss, both of which can increase HCO_3^- levels.
- Corticosteroid excess, as in Cushing's syndrome or hyperaldosteronism, can lead to both metabolic alkalosis and hypokalemia.

Expected compensation in metabolic alkalosis

Although metabolic alkalosis consistently causes hypoventilation, the timing and degree of this ventilatory compensation response to metabolic alkalosis is not always predictable, because other factors such as pain and hypoxemia can stimulate hyperventilation and overcome the hypoventilatory stimulus of metabolic alkalosis [8]. The maximal respiratory compensation by hypoventilation can elevate arterial pCO_2 to >80 mmHg (Table 36.3 and Fig. 36.2) [8].

In the initial phase of metabolic alkalosis, both pH and HCO_3^- are increased and pCO_2 remains normal. Within 2 hours, compensation by hypoventilation is expected, but maximal compensation requires 12–24 hours, which increases the blood pCO_2 to match the elevated HCO_3^-

and restore the HCO_3^-/pCO_2 ratio to normal. In general, the increase in pCO_2 from compensation by hypoventilation increases according to the HCO_3^- level in a predictable manner, as described in Table 36.3. If the actual pCO_2 deviates from this expected pCO_2 by $>$ 3−4 mmHg, the patient may have a mixed disorder due to either an underlying respiratory acidosis (if >3 mmHg higher) or alkalosis (if >3 mmHg lower).

Treatment of metabolic alkalosis

The optimal treatment for metabolic alkalosis is correcting the underlying cause. In some cases, simple hydration will gradually correct metabolic alkalosis if renal function is normal. If Cl^- depletion is present, Cl^- replacement by administration of NaCl or KCl also allows the kidney to excrete HCO_3^-. Hypokalemia is corrected, if present. Rarely, intravenous dilute HCl administration is necessary [11].

Respiratory acidosis

Respiratory acidosis is ventilatory failure that leads to a decreased pH and increased pCO_2. Acute respiratory acidosis occurs by either: decreased breathing or ventilation, increased production of CO_2 by the body, or excess CO_2 in the inspired gas [1,8]. Normally, any increased production of CO_2 promptly stimulates hyperventilation, which maintains arterial pCO_2 at normal levels. If arterial pCO_2 is increased, it almost always involves a problem with ventilation. Causes of respiratory acidosis include the following:

- Impaired pulmonary gas exchange by alveoli, such as with chronic bronchitis, emphysema, or asthma.
- Acute airway obstruction caused by aspiration or blockage of airway tube.
- Insufficient delivery of blood to the lungs from circulatory failure or pulmonary embolism.
- Impaired function of respiratory center caused by head trauma, drugs (opiates, barbiturates, etc.), toxins, or anesthesia.
- Weakness of respiratory muscles.
- Mechanical ventilation that does not provide either adequate oxygen or adequate removal of CO_2 [12].

Expected compensation in respiratory acidosis

During the acute phase, plasma buffering of the elevated CO_2 increases the HCO_3^- slightly, by approximately 1 mmol/L for each 10 mmHg rise in pCO_2 (Table 36.3 and Fig. 36.2). Over the next few hours, the kidneys increase reabsorption of HCO_3^-, which elevates serum HCO_3^- by about 2 mmol/L. As this continues into the chronic phase (over 24 hours), the HCO_3^- slowly rises and plateaus after 2−5 days with normalization of the pH. The expected blood level of HCO_3^- may be calculated, as described in Table 36.3. As an example,

* In chronic respiratory acidosis, if the pCO_2 has increased from 40 mmHg to a steady state of 60 mmHg over 3 days, the HCO_3^- should have risen to about 8 mmol/L. That is, if the patient's normal HCO_3^- was 24 mmol/L, it would be about 32 mmol/L after 3 days of metabolic compensation, with a normal pH.

Treatment of respiratory acidosis

Treatment should be directed at correcting the underlying cause: (a) restore adequate lung ventilation by such means as endotrachial intubation or positive pressure ventilation; (b) restore pulmonary circulation; or (c) correct any drug overdose. Care must be taken to avoid correction of the elevated pCO_2 too rapidly. Administration of bicarbonate is rarely done, because HCO_3^- crosses the blood−brain barrier slowly, which can elevate the blood pH without affecting the CNS pH [8].

Respiratory alkalosis

Respiratory alkalosis is caused by any condition that leads to hyperventilation. If a normal ventilatory rate of 12−15/min increases to $\geq$ 20/min, CO_2 will be lost faster than it is produced, leading to an increased pH and decreased pCO_2. Causes of respiratory alkalosis include the following:

- Hypoxemia-induced hyperventilation, as caused by breathing oxygen-poor air or exposure to high altitude.
- Pulmonary embolism or pulmonary edema, in which oxygen transport across the alveolar membrane is impaired to a greater extent than is CO_2 transport.
- Hyperventilation induced by anxiety, pain, or pregnancy.
- Excess mechanical ventilation, usually from aggressive use of the ventilator to increase arterial oxygen tension.
- Drugs such as salicylate, nicotine, or progesterone, which can cause hyperventilation.
- Central nervous system disorders resulting from conditions such as sepsis or trauma.

Expected compensation in respiratory alkalosis

Because metabolic compensation by the kidney to increase excretion of HCO_3^- is relatively slow, it is time-dependent and occurs in two phases, as described in Table 36.3 (see also Fig. 36.2), for example:

- Acute respiratory alkalosis (<24 hours): If pCO_2 decreases from 40 to 30 mmHg, this 10-mmHg fall should be associated with a pH increase of 0.08

(pH ~ 7.48) and bicarbonate should decrease by about 2 to ~22 mmol/L.
- Chronic respiratory alkalosis (2–5 days): If the pCO_2 is 30 mmHg after 2 days of hyperventilation, the pH should be increased by 0.03 (pH ~7.43) and HCO_3^- should be decreased by about 4 to ~20 mmol/L.

Note that the lower limit for metabolic compensation in hyperventilation is a plasma HCO_3^- of ~12–14 mmol/L. If plasma HCO_3^- is <12 mmol/L, an underlying metabolic acidosis should be suspected.

Treatment of respiratory alkalosis

Treatment is aimed at correcting the underlying condition. Most urgently, hypoxemia must be corrected by giving oxygen. Salicylate or other drug overdoses should be treated appropriately. Any anxiety-induced respiratory alkalosis may simply be treated with reassurance or rebreathing expired air from a paper bag (higher CO_2 content) that can be helpful in otherwise healthy persons [8].

Detecting mixed acid–base disorders

Mixed acid–base disorders occur when multiple primary acid–base disorders occur at the same time. These are common in hospital populations, especially in the emergency department (ED) and in critical care. Although there are equations and diagrams that can aid in the diagnosis, these relationships will not hold in patients with complex acid–base disorders, especially if they occur and/or resolve at different times. There is no substitute for a careful review of a patient's clinical course and serial blood gas measurements.

The importance of identifying mixed acid–base disorders lies in their diagnostic and therapeutic implications. For example, the development of a primary metabolic alkalosis (increased [HCO_3^-]) in a patient with chronic obstructive airway disease who is being treated with diuretics should alert the clinician to possible potassium depletion. In another example, a patient presenting with a mixed respiratory alkalosis and metabolic acidosis should be evaluated for salicylate intoxication.

Does the expected compensation occur?

A simple concept to remember in a primary acid–base disorder is that if the expected compensation does not occur within an expected time frame, a mixed disorder should be suspected. For example, in a primary metabolic acidosis, the lungs are expected to compensate this excess metabolic acid by hyperventilating (a respiratory alkalotic process) to remove the respiratory acid CO_2 and return the pH toward normal. If this does not occur in the time or to the degree expected, the person is considered to have an underlying respiratory acidosis. Refer to the examples below and to Fig. 36.2 and Table 36.3 for the expected responses. *Note that these equations should only be used as guides and not as absolute diagnoses, especially when a patient has multiple acid–base disorders that have occurred over time.*

Metabolic acidosis

The body's natural response during the first 12–24 hours is hyperventilation (a respiratory alkalotic process), which decreases the blood pCO_2 in a somewhat predictable way and increases pH back toward normal.

Metabolic alkalosis

The response to metabolic alkalosis is respiratory hypoventilation (a respiratory acidotic process), which increases the pCO_2 and decreases pH back toward normal.

Respiratory acidosis or respiratory alkalosis

If a patient has a primary respiratory disorder, either acidosis or alkalosis, simple rules predict the expected pH changes versus the change in pCO_2 during the acute and chronic phases. During the acute phase of respiratory disorders, for each 10-mmHg rise or fall in pCO_2, the pH should decrease or increase (respectively), by 0.08. As this progresses into the chronic phase (1–2 days), for each 10-mmHg rise or fall in $pCO2$, the pH should have changed by only 0.03 [1,8,10]. Fig. 36.2 illustrates the relationships among pCO_2, pH, and HCO_3^- for acute and chronic respiratory changes. If the blood gas data clearly do not reflect these temporal relationships, it suggests an additional disorder is present and that the patient has a mixed disorder.

Delta ratio or delta gap: does the change in anion gap match the change in bicarbonate?

The delta ratio (DR) or the delta gap (DG) is the calculation that may be useful in a patient with metabolic acidosis to indicate if a mixed acid–base disorder may be present. Typically, in metabolic acidosis, an increase in AG from 12 mmol/L and a decrease in HCO_3 from 24 mmol/L should approximately equal each other. This is called the "DR" [8], calculated as:

$$\text{Delta ratio} = \frac{\text{AG} - 12}{24 - \text{HCO}_3} \tag{36.x}$$

Obviously, for these calculations to be useful, there must be a significant change in the AG from 12; otherwise, the DR will be zero. In a typical AG metabolic acidosis, this "DR" would be 1.0. If the DR is greater than 1,

TABLE 36.4 Examples of delta ratio and delta gap calculations for metabolic acidosis.

pH	Na	Cl	HCO_3^-	AG	DR	DG	Condition
7.40	140	104	24	12	0/0	0	Normal: no metabolic acidosis
7.25	140	104	12	24	12/12 = 1.0	0	High AG metabolic acidosis
7.15	140	110	14	16	4/10 = 0.4	− 6	High AG metabolic acidosis + normal AG metabolic acidosis
7.30	140	98	22	20	8/2 = 4.0	+ 6	High AG metabolic acidosis + metabolic alkalosis

AG, Anion gap; *DG*, delta gap; *DR*, delta ratio.

it suggests there is more HCO_3^- in the blood than expected from the increased AG, so that an additional metabolic alkalosis may be present. If the DR is less than 1, it suggests less HCO_3^- is present than expected from the increased AG, so that an additional non-AG metabolic acidosis, such as hyperchloremic metabolic acidosis, may be present.

This may also be calculated as a DG as follows:

$$DG = (AG - 12) - (24 - HCO_3)$$

$$DG = AG + HCO_3 - 36$$

Since the AG is Na−Cl−HCO_3, the DG calculation can be further simplified as follows:

$$DG = Na - Cl - 36$$

An elevated AG with a DG $\geq$ 6 suggests a mixed AG metabolic acidosis with a metabolic alkalosis, because the fall in HCO_3^- is less that the rise in AG. An elevated AG with a significant negative DG (≤ -6) suggests a mixed disorder: a high AG metabolic acidosis with a hyperchloremic (normal AG) metabolic acidosis is present, because the fall in HCO_3^- is greater than the rise in AG [13]. Table 36.4 gives some examples of DR and DG calculations for various metabolic acidosis.

Tips for diagnosing mixed acid−base disorders

The following are additional tips to diagnose mixed acid−base disorders:

- If the pH is well within normal and both the HCO_3^- and pCO_2 are abnormal, this strongly suggests a mixed disorder.
- Whenever HCO_3^- and pCO_2 are abnormal in the opposite direction, the compensatory response should always be in the same direction as the change caused by the primary disorder.
- Simple acid−base disorders do not overcompensate, that is, compensation will not cause an acidemia to become an alkalemia.

Evaluating blood gas results

Reference and critical ranges

The reference ranges for both arterial and venous blood are shown in Table 36.1 for the common blood gas and acid−base parameters [3,9], along with typical critical limits for each parameter. Notes:

- Venous pH and pCO_2 may be used to assess acid−base and ventilatory status, with venous pH approximately 0.02−0.04 units lower than arterial pH and venous pCO_2 approximately 3−7 mmHg higher than arterial values.
- Venous pO_2 has no clinical value and cannot be used to approximate arterial pO_2.
- The pO_2 of arterial blood varies with age, decreasing by about 1 mmHg with every year over 60 years of age.

Evaluating the acid−base status

Ideally, evaluating acid−base disorders, including mixed acid−base disorders, would include simultaneous information on the patient's clinical history, the electrolyte results, and the blood gas results. With modern blood gas analyzers, the blood gas results are often available with the electrolyte results.

The acid−base and blood gas results (pH, pCO_2, pO_2, and HCO_3^-) may be evaluated in several steps, as follows:

1. Evaluate the patient's clinical history and current status to anticipate conditions associated with acid−base disorders.
2. Evaluate the pH.
3. Evaluate the ventilatory (pCO_2) and metabolic (HCO_3^-) status to determine the primary disorder.
4. Evaluate laboratory and clinical data for a possible mixed disorder:
   a. For a primary metabolic (or respiratory) disturbance, is respiratory (or metabolic) compensation appropriate?

TABLE 36.5 Examples of the expected acid–base disorder associated with various clinical conditions.

Clinical condition	Expected acid–base disorder
Cardiac arrest	Metabolic acidosis
Pulmonary arrest	Respiratory acidosis
Hyperventilation (many causes)	Respiratory alkalosis
Blockage of endotrachial tube ⇒ (after removal of blockage from tube)	Respiratory acidosis ⇒ (respiratory alkalosis)
Congestive heart failure leading to hyperventilation	Respiratory alkalosis
Vomiting	Metabolic alkalosis
Diarrhea	Metabolic acidosis
Shock state (inadequate perfusion)	Metabolic acidosis
Pulmonary edema	Respiratory alkalosis (hypoxia leads to hyperventilation)
Severe pulmonary edema	Respiratory acidosis
Diuretic therapy	Metabolic alkalosis
Drug intoxication	Respiratory acidosis (respiratory arrest)
Bicarbonate therapy	Metabolic alkalosis
Poor perfusion	Metabolic acidosis

b. Do any other parameters, such as electrolytes, lactate, AG, or DR indicate that another acid–base disturbance is present?
c. Is the patient's history consistent with the blood gas results?

The evaluation of oxygen status is mostly independent of the acid–base interpretation, so it will be evaluated in a separate section.

Step 1: evaluate the patient's status and history to anticipate possible acid–base abnormalities

Many conditions are associated with acid–base disorders, as listed in Table 36.5. A perceptive clinical evaluation can be a most important part of evaluating a patient for acid–base disorders.

Step 2: evaluate the pH

An abnormal pH indicates that an acidosis or alkalosis has occurred and the extent of the acid–base disorder, and it may suggest that compensation has occurred. However, the pH by itself does not indicate whether a mixed disorder is present. Consider the following examples:

- pH 7.20 confirms that a severe acidosis is present and that compensation is either in its early stages or ineffective in controlling the acidosis. Further investigation is required to determine the metabolic or respiratory origin of the acidosis and whether more than one acidotic process is present.
- pH 7.48 indicates that a mild alkalosis is present. Further information is required to determine the cause of the alkalosis and whether the alkalosis is in its early stages (and may get worse), has been nearly compensated, or is part of a mixed disorder.
- A normal pH may indicate that the patient has no acid–base disorder, or a patient may have both acidotic and alkalotic events (primary or compensatory) that have offset each other. The patient's history and the HCO_3^- and pCO_2 should be considered.

Step 3: evaluate the ventilatory and metabolic statuses

The simplest approach to evaluating the pCO_2 and HCO_3^- is to consider each parameter separately as indicating acidosis, alkalosis, or normal status, then evaluate them along with pH to determine whether the primary disorder is respiratory or metabolic and whether a mixed disorder or compensation is present.

A decreased pCO_2 indicates a respiratory alkalotic process, which may be either primary or compensatory.

An increased pCO_2 indicates a respiratory acidotic process, either primary or compensatory. A decreased HCO_3^- indicates a metabolic acidosis, either primary or compensatory. An increased HCO_3^- indicates a metabolic alkalosis, either primary or compensatory.

Note: The terms "respiratory alkalosis," "metabolic acidosis," etc., refer to pathologic processes and not to compensation. However, to simplify interpretation, momentarily regard any decreased pCO_2, for example, as a respiratory alkalotic process. It will become clear whether it is a primary abnormality or appropriate compensation as other factors are considered.

A normal pCO_2 and HCO_3^- indicate normal ventilatory and metabolic status. However, if one of these parameters is normal when the other is abnormal, then a mixed disorder may be present. That is, a lack of an appropriate compensatory response suggests an additional disorder is present.

Examples

A pCO_2 result of 25 mmHg indicates that a respiratory alkalotic process is present. However, referring to Fig. 36.2, it could be an acute or chronic respiratory alkalosis, or even a compensated metabolic acidosis. It depends on the HCO_3^- and pH associated with this pCO_2 to determine whether it is a primary abnormality or appropriate compensation. If the HCO_3^- is 14 mmol/L (with a pH of ~7.32), then the pCO_2 of 25 mmHg probably indicates compensating hyperventilation (a respiratory alkalotic process) for a primary metabolic acidosis.

In another example for a pCO_2 of 25 mmHg, if the HCO_3^- is 23 mmol/L and the pH is ~7.53, this suggests acute hyperventilation, but this diagnosis would also depend on the duration of hyperventilation. If acute, the metabolic compensatory response to eliminate HCO_3^- is progressing about as expected. However, if the hyperventilation is chronic, then the higher than expected HCO_3^- suggests that, because the kidney is unable to compensate adequately by eliminating HCO_3^-, a metabolic alkalosis may also be present.

Steps 4a–4c: evaluate for a possible mixed disorder

4a. Is the compensation adequate for the primary disorder?

To evaluate whether or not the expected compensation is occurring for the primary acid–base disorder requires serial measurements of blood gas and electrolyte results; refer to Fig. 36.2 and Table 36.3 and the "Expected Compensation" descriptions in the sections "*Clinical Abnormalities of Acid-Base Balance*" and "*Mixed Acid-Base Disorders.*"

4b. Do other laboratory results suggest an additional acid–base disorder is present?

Several common laboratory tests may help determine whether additional acid–base disorders are present [8,10,12].

Potassium

Hypokalemia is frequently associated with a metabolic alkalosis.

pH

A normal pH combined with abnormal results for both HCO_3^- and pCO_2 warrants consideration of a mixed acid–base disorder.

Chloride

Hyperchloremia results from HCO_3^- loss, either from a primary metabolic acidosis as with renal loss of HCO_3^- with retention of Cl, or from the metabolic acidotic compensation for respiratory alkalosis. Hypochloremia may cause metabolic alkalosis by promoting renal reabsorption of HCO_3^- or may be caused by compensation for chronic respiratory acidosis (as HCO_3^- is retained and Cl^- is excreted).

Anion gap

An elevated AG $[(Na^+) - (Cl^- + HCO_3^-)]$ of >20 mmol/L indicates metabolic acidosis. AG is also useful in differentiating the causes of metabolic acidosis, as shown in Table 36.2 [2,13].

While often very useful, AG is calculated from three or four measurements and is subject to variation from these measurements. As an example, some patients with a clearly elevated blood lactate will not have an elevated AG [2,13].

Delta ratio

The "DR" should also be evaluated, as discussed earlier in the section "Detecting mixed acid–base disorders." Since this is based on the AG, it is also subject to analytical variations in the methods and also variation in a specific patient's "normal" AG.

Lactate

A rising blood lactate is a sensitive indicator of metabolic acidosis resulting from poor oxygenation of tissues or of mitochondrial dysfunction.

Creatinine

An elevated creatinine indicates renal insufficiency and possible uremic metabolic acidosis.

4c. Does the patient have other conditions associated with an acid–base disorder?

An evaluation of blood gas results must always include the patient's historical and physical findings. Many conditions are frequently associated with an acid–base disorder, with some common ones listed in Table 36.5, especially if the blood gas results cannot readily be explained by the patient's "primary" disorder. The presence of more than one such condition should alert the clinician to the possibility of a mixed acid–base disorder. Mixed acid–base conditions are quite common.

Evaluating oxygen status

Hemoglobin binding

Hb binding to O_2

Hb consists of four heme molecules attached to four globin molecules and has the essential abilities to (1) bind oxygen in the lungs then transport and release oxygen to the tissues, and (2) transport H^+ and carbon dioxide from the tissues to the lungs. Each of the four heme groups contains an Fe ion and can bind one molecule of oxygen. Structural changes occur with oxygen binding that results in color changes to the molecule [9].

Several factors affect Hb binding of oxygen. In addition to pO_2, these factors are H^+, temperature, pCO_2, and 2,3-diphosphoglycerate (2,3-DPG). The effects of these factors are easy to remember, because they are inversely related to oxygen binding. That is, an increase in H^+ concentration, temperature, carbon dioxide pressure, or 2,3-DPG concentration will decrease the affinity of Hb for oxygen. In the tissues, blood enters an area of relative warmth, acidity, elevated pCO_2, and low pO_2, conditions that promote loss of O_2 to tissues and binding of H^+ by Hb.

The role of 2,3 DPG on Hb binding and release of O_2 is more complex. 2,3 DPG is a molecule largely contained in RBCs that "cooperates" with O_2 to control when Hb binds or releases O_2. In the lungs with a high pO_2, oxygen binds to Hb, which promotes the release of 2,3-DPG. This release of 2,3-DPG further increases the affinity of Hb for O_2 so that Hb binds more O_2. The loss of 2,3-DPG favors a Hb conformation (R form) with a higher affinity for O_2 [14]. As blood enters the tissues, the increased H^+ ions and pCO_2 promote release of O_2, which then favors binding of 2,3-DPG to Hb. This causes a conformational change to the T form of Hb that has a lower affinity for O_2, therefore leading to further release of oxygen to the tissues. The sigmoidal relationship between pO_2 and sO_2 is well known as the Hb-oxygen dissociation curve, shown in Fig. 36.3.

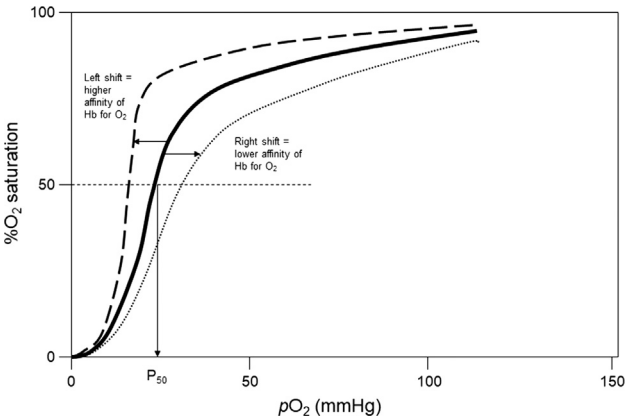


FIGURE 36.3 Oxyhemoglobin dissociation curve for whole blood. The solid middle curve represents the percent of hemoglobin that becomes saturated with oxygen as the oxygen pressure (pO_2) is increased. The dashed curve represents a *left shift* of the curve when the affinity of Hb for oxygen is increased by lower temperature, lower H ion concentration, lower oxygen pressure, etc. The dotted curve represents a *right shift* of the curve when the affinity of Hb for oxygen is decreased by higher temperature, higher H ion concentration, higher oxygen pressure, etc. The P_{50} is the oxygen pressure that gives Hb a 50% saturation and is an index of the overall affinity of Hb for oxygen. A *left shift* of the curve decreases the P_{50} and a *right shift* increases the P_{50}.

Fetal Hb, a form of Hb present in decreasing amounts during the first month of life, does not bind 2,3-DPG as well as adult Hb, so fetal Hb tends to stay in a conformation with a higher affinity for O_2 than does adult Hb. This is essential so that the fetal blood can take O_2 away from maternal blood. Fetal Hb also has a slightly altered absorbance spectrum, which can falsely elevate readings for COHb. Different wavelengths and various correction factors are used in the analyzers to minimize this interference.

Nitric oxide (NO) also plays a role in DO_2. NO regulates vascular tone, which can affect both blood perfusion and DO_2 to tissues. Being structurally very similar to molecular oxygen (O_2), NO also binds to the heme iron of Hb and can form both metHb and S-nitroso Hb. The binding and release of oxygen by Hb may selectively release NO to the cell membrane, where it affects blood pressure [15].

Disorders of oxygenation: hypoxemia and tissue hypoxia

Utilization of oxygen requires many processes to function smoothly:

- air intake by breathing or ventilation;
- air entry into the alveoli;
- pulmonary gas exchange with functional alveoli;
- oxygenation of blood and Hb;

- O_2 transport to tissues and cells, with adequate Hb and systemic blood flow;
- O_2 release and diffusion into cells; and
- O_2 utilization by the mitochondria to produce adequate ATP and so on.

Inadequate oxygenation of blood may be caused by:

- low *FI* O_2, such as from inadequate oxygen in inhaled anesthetics, fires, or high altitudes;
- hypoventilation leading to insufficient O_2 in alveolar air;
- mismatch of pulmonary ventilation to pulmonary perfusion: ventilation/perfusion (V/Q) mismatch;
- diffusion (gas exchange) impairment between alveolar air and blood: pulmonary embolism, emphysema, and pulmonary edema;
- intrapulmonary shunting: blood goes from the right heart to the left heart without contacting functional alveoli as with pneumonia; and
- anemia: insufficient O_2 carrying capacity in blood.

Poor tissue oxygenation or O_2 utilization may be caused by the following:

- inadequate blood flow to tissues due to diminished cardiac output or altered blood flow;
- inadequate uptake of O_2 by the tissues; and
- mitochondrial dysfunction

Both pO_2 and blood lactate may be used to evaluate these conditions.

O_2 delivery to tissues

Gas exchange in the lungs

The pO_2 of atmospheric air (21% O_2) is $\sim$150 mmHg, accounting for the contribution of water vapor (45 mmHg) to the atmospheric pressure: $(760-45) \times 0.21 = 150$ mmHg. An adult breathing air with a pO_2 of 150 mmHg should have an alveolar pO_2 $\sim$100 mmHg because of CO_2 in the alveolar air and an arterial pO_2 of 80–100 mmHg. The reference range for arterial pO_2 also depends on age [9].

Diminished gas exchange in the lungs may be caused by:

- pulmonary congestion caused by asthma, edema, or inflammation;
- fibrosis;
- alveolar dead space;
- intrapulmonary shunting;
- mucus hypersecretion (bronchitis);
- loss of alveolar compliance; or
- insufficient surfactant (neonatal respiratory distress syndrome).

Alveolar dead space refers to the gas volume in lungs that does not undergo gas exchange and is caused by unperfused alveoli. Dead space is evaluated by comparing the end-tidal pCO_2 with the arterial pCO_2, where a lower than expected end-tidal pCO_2 indicates a possible increased dead space. In intrapulmonary shunting, venous blood does not get oxygenated, because some blood passes through the pulmonary capillaries of nonfunctioning alveoli. When blood from these "shunts" mixes with oxygenated blood in the pulmonary vein, blood of a lower pO_2 and oxygen content is delivered to the systemic circulation [9]. Patients with excess pulmonary shunting require more oxygen than expected.

Blood flow to tissues

Although normal blood flow provides sufficient oxygen and nutrients to all organs and tissues, abnormalities in flow may develop that cause inadequate perfusion of general or localized areas of the body. Such conditions may lead to diminished oxygen content in arterial blood, shock states, tissue necrosis, and possibly death. Some common causes of poor perfusion are:

- inadequate cardiac output;
- decreased blood volume (hypovolemia);
- emboli;
- vasoconstriction;
- sepsis.

Evaluation of oxygenation status and arterial oxygenation

The blood gas parameters often used for assessing arterial oxygenation are as follows:

- the arterial oxygen tension (pO_2), especially in relation to the oxygen content of the air the patient is breathing (FIO_2);
- the sO_2 or the $\%O_2$-Hb in blood;
- the Hb concentration in blood.

The following parameters also may be calculated from the measured blood gas results:

- the oxygen content in blood;
- alveolar–arterial pO_2 difference or gradient;
- V/Q mismatch and intrapulmonary shunting of blood, in which blood flows into the lung but does not perfuse functional alveoli; and
- tissue oxygenation, assessed by blood lactate measurement.

Arterial O_2 tension and sO_2 of Hb

The pO_2 and sO_2 values often are used for assessing arterial oxygenation, with the pO_2 being a more linear reflection of the efficiency of blood oxygenation by the lungs,

and the sO_2 (along with Hb) being a better indicator of the O_2 content of blood. Hypoxemia is defined as an arterial pO_2 that is below the minimally acceptable limit for persons of a given age group breathing room air [9]. As pO_2 decreases <60 mmHg, the degree of hypoxemia intensifies, with severe hypoxemia occurring below 40 mmHg.

Ideally, the diagnosis of hypoxemia is made when the patient is breathing room air. However, if the patient is receiving oxygen therapy, changing the patient to breathing room air is likely not an option. So for a patient breathing oxygen-enriched air, the likelihood of hypoxemia may be predicted by comparing the arterial pO_2 with the FIO_2 (percent oxygen being breathed). If the arterial pO_2 (mmHg) is less than five times the FIO_2 (%), the patient's alveolar ventilation may be inadequate and the patient would probably be hypoxemic if breathing room air [9].

p_aO_2/FIO_2 ratio

The ratio of the p_aO_2 to FIO_2 is a simple, readily available clinical calculation used as one of the criteria to determine the presence and severity of the acute respiratory distress syndrome (ARDS). A higher ratio indicates better lung function or less severe disease [16]; p_aO_2/FIO_2 ratios follow:

- $\sim$350–500 = expected normal (lower values expected with aging);
- 201–300 = mild ARDS;
- 101–200 = moderate ARDS; and
- <100 = severe ARDS.

ARDS is a disorder characterized by diffuse damage to the alveolar–capillary membrane leading to noncardiogenic pulmonary edema and hypoxemia, caused by a number of insults, both pulmonary and nonpulmonary. The most common causes of ARDS include pneumonia, sepsis, aspiration, trauma, and pancreatitis, which lead to a systemic inflammatory response and injury to the alveolar–capillary interface. This causes exudation of protein-rich fluid into the alveoli and deficiency of surfactant, leading to alveolar collapse (atelectasis). ARDS is characterized by an acute onset of hypoxemic respiratory failure with a p_aO_2/FIO_2 ratio less than 300.

Hb concentration

Because a normal Hb concentration in blood (12–17 g/dL depending on gender) provides excess oxygen carrying capacity, a slightly decreased Hb of 10 g/dL is usually of no concern in critical care. Clinicians usually want to be sure the person is not severely anemic.

Blood O_2 content

Blood oxygen content is calculated from measurements of blood Hb and sO_2. Anemia will markedly lower the O_2 content, as will hypoxemia (low pO_2/sO2), hypercapnia (high pCO_2), acidemia (low pH), and hyperthermia, all of which cause Hb to be less saturated for a given pO_2.

Alveolar–arterial oxygen difference or gradient (A–a)

The alveolar–arterial pO_2 difference (A–a) is a measure of the pO_2 difference between the alveolar air and the arterial blood [10,17,18]. It may be useful for determining both the cause (i.e., ventilatory failure vs. oxygenation failure) and the severity of hypoxemia [19]. The alveolar pO_2 (p_AO_2) may be calculated as the difference between the atmospheric pO_2 of 150 mmHg [$(760-45)\times 0.21 = 150$ mmHg] (45 mmHg is atmospheric water vapor pressure) and the blood pCO_2 divided by a gas exchange factor (0.8) that represents the ratio between the arterial and alveolar pCO_2. Thus for typical conditions in a healthy person with measured arterial pCO_2 of 40 mmHg and arterial pO_2 of 95 mmHg (paO_2), the alveolar pCO_2 would be 50 (40/0.8) and the alveolar–arterial pO_2 difference would be:

$$p_AO_2 - paO_2 = \left(150 - 40/0.8\right) - 95 = 5 \text{ mmHg}.$$

- A low paO_2 with a significantly larger A–a difference, suggests oxygenation failure by poor alveolar gas exchange, caused by edema, inflammation, fibrosis, or increased pulmonary shunting.
- A hypoxemic patient with a normal A-a difference implies that the lungs are not ventilating enough air (hypoventilation).

V/Q mismatch and intrapulmonary shunting

The term V/Q refers to the amount of air entering the alveoli per minute (V) relative to the capillary perfusion (Q) of those alveoli.

- A V/Q ratio of 1.0 means that an amount of ventilation in an alveolar unit (such as 1 mL/min) is available to exchange gases with an equal amount of alveolar blood (1 mL/min).
- A V/Q ratio of 2.0 indicates twice as much alveolar ventilation for the amount of alveolar capillary perfusion.
- A ratio of 0.5 indicates half as much ventilation as perfusion [17].

A simple example of a condition leading to V/Q mismatch (V/Q > 1) would be a pulmonary embolism where a segment of the lung receives ventilation, but the blood flow to that ventilated area is blocked by a blood clot. Other conditions that can lead to V/Q mismatching include emphysema and pulmonary edema.

Intrapulmonary shunting is a condition of V/Q mismatching with a V/Q ratio of 0 for a given area of the lung leading to an overall V/Q < 1. A shunt indicates that blood flows through a portion of the lungs without coming in contact with functioning alveoli. This can occur in the setting of atelectasis, pneumonia, or complete bronchial obstruction where pulmonary blood perfuses nonfunctional or nonventilated alveoli.

The percent of shunted blood is the ratio of shunted cardiac output to total cardiac output, which may be estimated from O_2 content measurements of both arterial and venous blood. The following equation is an approximate calculation, assuming that blood in the pulmonary capillaries is 100% saturated:

$$\%\ \text{Shunt fraction} = \frac{100 - sO_{2\text{art}}}{100 - sO_{2\text{ven}}}$$

As an example, for a patient with an sO_{2art} of 92 and an sO_{2ven} of 70%, the % shunt fraction would be 27%.

Evaluating tissue oxygenation

Tissue oxygenation can be assessed from the calculations of DO_2 and VO_2, which require measurements of blood oxygen content and cardiac output. More simply, measurement of blood lactate is a laboratory test that provides a sensitive index of the overall state of DO_2 and oxygen utilization by mitochondria.

Collection and handling of samples for blood gas analysis

Use of arterial versus venous blood for blood gas and acid–base measurements

Perhaps more than any other analytes, pO_2, pCO_2, and to some extent pH change markedly from arterial to venous blood. Arterial blood is mostly preferred over venous blood for blood gas analysis of pH, pCO_2, and pO_2, because arterial pO_2 indicates the ability of the lungs to oxygenate the blood with alveolar air, and arterial blood provides an index of the oxygen and nutrients that will be provided to the tissues and cells.

Because typical venous blood collected from an arm vein represents oxygen metabolism only in the arm, venous blood representing oxygen metabolism of the entire body is sometimes needed. This requires that "mixed" venous blood be collected, which exists in the pulmonary artery and can be collected with a Swan–Ganz catheter [9]. Both mixed venous and arterial blood are needed when determining parameters such as P50 and VO_2.

Collection and handling of blood

Blood collected for blood gas analysis is highly susceptible to changes in pO_2. Anaerobic conditions during collection and handling are essential, because room air has a pCO_2 of nearly 0 and a pO_2 of ~150 mmHg. The factors that must be controlled are:

- removal of all air bubbles;
- use of the proper anticoagulant;
- appropriate use of plastic syringes (glass syringes rarely used);
- the temperature of storage before analysis;
  - agitation of the blood specimen; and
- the length of delay between collection and analysis of blood.

The complete removal of all air bubbles is especially important before sending blood in a syringe by pneumatic transport, which can markedly affect pO_2 by up to 25 mmHg, depending on the initial pO_2 [20,21]. pCO_2 is almost unaffected by air bubbles.

Although liquid heparin at <10% (vol./vol.) of the volume of blood should have a small effect on pH, pCO_2, or pO_2, the use of liquid heparin will dilute other constituents in blood, such as electrolytes and glucose that are typically analyzed simultaneously in current blood gas and electrolyte analyzers. Therefore only dry heparin should be used as an anticoagulant. While plastic syringes are used for nearly all blood gas measurements, they have a potential disadvantage due to their ability to absorb oxygen [22]. When stored in ice water, because of the increased oxygen affinity of Hb at cold temperatures, blood can absorb oxygen dissolved within the wall of the syringe and transmitted through the plastic. This effect is most pronounced in samples with a pO_2 of ~100 mmHg and above, that is, when Hb is already nearly fully saturated with oxygen and is unable to buffer any added O_2. A pO_2 of 100 mmHg may increase by 8 mmHg during 30 minutes of storage on ice.

When Hb is less saturated (for example, at a pO_2 of 60 mmHg), it is better able to buffer the additional oxygen, with little measurable change in pO_2. Because glass is not permeable to O_2, pO_2 is not affected in blood stored in glass syringes in ice.

Storage of blood specimen

In general, storage of blood in plastic syringes at room temperature is acceptable if analysis is within 15 minutes

[22]. When blood is stored at ambient temperatures (22°C–24°C; 72–75 °F) for >30 minutes, pO_2 has been reported to both increase and decrease, with the increase likely due to oxygen permeability of plastic syringes [22–24]. Storage for 30 minutes at ambient temperature changed pO_2 by about 2 mmHg at a pO_2 of 100 mmHg [22]. In most samples at ambient temperature, pCO_2 changes <1 mmHg, and pH changes <0.01 unit. However, samples from patients with extreme leukocytosis can dramatically change pH, pCO_2, and pO_2 (and glucose and lactate) when stored at room temperature [10]. These samples must be analyzed as soon as possible.

Cord blood gases

Blood gas values on blood collected from the umbilical cord at birth can provide valuable data for assessing the respiratory and metabolic status of the newborn. They are more objective than apgar scores at identifying hypoxia and acidosis in the neonate that may occur from asphyxia, respiratory distress, or other conditions. Note that umbilical venous blood reflects the combined effects of maternal acid–base status and placental function, while umbilical arterial blood represents the condition of the neonate. Therefore, in most cases, blood from the umbilical artery should be collected, because it contains blood returning from the fetus to the placenta and more accurately represents the neonatal acid–base status. Some studies have concluded that collection of both umbilical arterial and venous blood provides clinically useful information on the pathogenesis of acidosis [25,26].

When the fetus cannot eliminate sufficient CO_2 via the placenta, the pCO_2 increases leading to a "respiratory" acidosis. If O_2 exchange is inadequate (from multiple causes), a metabolic acidosis results. Some critical results for cord blood are:

- pH < 7.0 in arterial cord blood.
- pCO_2 difference between arterial cord blood and venous cord blood > 25 mmHg.

The relevance of pO_2 results on cord blood is not clearly established. Reference ranges for both arterial cord blood and venous cord blood have been published [25] and are summarized in Table 36.6.

Temperature correction of blood gas results

Whether to temperature-correct, or not correct, blood gas results appear to be a never-ending controversy, and the reader is referred to an excellent recent discussion on this issue [27]. While patients may have pathologic or environmental hyperthermia or hypothermia, the main issue is with patients who are deliberately cooled (therapeutic hypothermia) to increase survival in situations such as cardiac surgery, myocardial infarction, ischemic stroke, traumatic brain injury, or sepsis. With hypothermia, pCO_2 is decreased which increases the pH. The effects of temperature on pO_2 are more complex, but hypothermia by itself will lower the pO_2. The controversy is whether to temperature-corrected blood gas measurements back to the patient's body temperature (such as 30°C), which is called the "pH-stat" hypothesis, or interpret the results by the analyzers at 37°C, which is the "alpha-stat" hypothesis. This report cited other reports to conclude that for adult patients made hypothermic during cardiac surgery, uncorrected blood gas results are more appropriate for patient management, while for similar pediatric patients, corrected blood gas results appear more appropriate.

Usefulness of other tests on modern blood gas analyzer menus

Lactate

Blood lactate measurements have dramatically increased over the last 15 years, largely because they are frequently used in critically ill patients (notably in sepsis) to evaluate the initial condition of the patient and to monitor the effectiveness of therapy. In addition to the traditional view that elevated blood lactate is caused by an oxygen deficit to tissues, an elevated lactate can also be caused by mitochondrial dysfunction. This includes cyanide poisoning and sepsis, where several factors associated with sepsis inhibit mitochondrial function and elevate blood

TABLE 36.6 Reference ranges for both arterial and venous cord blood.

Analyte	Arterial cord blood reference range	Venous cord blood reference range
pH	7.14–7.42	7.22–7.44
pCO_2 (mmHg)	34–78	30–63
pO_2 (mmHg)	3–40	12–43

pO_2, Oxygen pressure.

lactate: inflammation, cytokines, platelet and endothelial activators, tissue necrosis factor, etc. Some common uses of blood lactate measurements include the following:

- In neonates during and after open-heart surgery for congenital heart disease.
- To evaluate patients who may require ECMO and for monitoring their progress.
- For triage in an ED setting to determine which patients need immediate care and for monitoring the effectiveness of therapy.
- In trauma patients, early identification of increased blood lactate followed by aggressive resuscitation improves survival. Survival was very high (98%–100%) in patients whose blood lactate normalized within 24 hours; was 75%–80% in patients whose blood lactate normalized in 24–48 hours; but was relatively poor in patients whose blood lactate could not be normalized by 48 hours [28].
- During open-heart surgery, patients are cooled to reduce VO_2. However, this can also cause perfusion abnormalities (vasoconstriction and shunting) that can lead to tissue hypoxia. Along with the anesthesia and drugs, these can all cause problems with oxygen metabolism and increase blood lactate. Thus, blood lactate measurements have become a means to monitor such patients.
- To identify high-risk ICU patients needing more aggressive therapy: For patients who responded to treatment by decreasing their lactate to less than 1.0 mmol/L, mortality was less than 7%.
- In goal-directed therapy for sepsis, blood lactate measurements have become an essential component of detecting high-risk patients (lactate >4.0 mmol/L), detecting hypodynamic state of low DO_2, and other lactate elevations due to inflammation, etc., but without evidence of tissue hypoxia [29]. As noted earlier, several factors associated with sepsis can elevate lactate, including inflammation, cytokines, platelet and endothelial activators, and tissue necrosis factor.

Glucose

The measurement of glucose in surgery and critical care is important as a means to maintain blood glucose levels within a range of about 140–180 mg/dL [30]. This more conventional range contrasts with the tight glycemic control with insulin infusion that has now been associated with higher rate of hypoglycemia and mortality [31]. A number of events unrelated to diabetes can also hinder the regulation of blood glucose levels in critically ill patients: infection and inflammation, fever, surgical trauma, general medical stress, glucocorticoid therapy, physical inactivity, etc. Hyperglycemia is also associated with undesirable effects, such as damage to certain brain and neuronal cells. While the optimal level of blood glucose to achieve is debatable, rapid measurements are necessary to avoid either hyperglycemia or hypoglycemia.

Met-Hb

Met-Hb may be used as a monitor or guide for appropriate use of inhaled NO which is used as a therapy of pulmonary hypertension in children after cardiac surgery. NO binds to Hb with great affinity and forms metHb by oxidation in the erythrocyte. Because the amount of functional Hb is thereby reduced, the patient may be more susceptible to anemia, acidosis, respiratory compromise, or cardiac problems. Typically, the proportion of NO in the inhaled gas is 5–40 ppm, which results in a maximum increase in metHb of 4% [32].

References

[1] A.W. Grogono, Acid-base tutorial: terminology. <http://www.acid-base.com> (accessed December 2008).

[2] J.A. Kraut, N.E. Madias, Serum anion gap: its uses and limitations in clinical medicine, Clin. J. Am. Soc. Nephrol. 2 (2007) 162–174.

[3] C. Higgins, Clinical aspects of the anion gap. <http://acutecare-testing.org> (accessed July 2009).

[4] F.D. Masevicius, A. Dubin, Has Stewart approach improved our ability to diagnose acid-base disorders in critically ill patients? World J. Crit. Care Med. 4 (1) (2015) 62–70.

[5] M.G. Scott, V.A. LeGrys, J.S. Klutts, Electrolytes and blood gases, in: C.A. Burtis, E.R. Ashwood, D.E. Bruns (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, fourth ed., Elsevier Saunders, St Louis, MO, 2006, pp. 983–1018.

[6] R.A. Robergs, f Ghiasvand, D. Parker, Biochemistry of exercise-induced metabolic acidosis, Am. J. Physiol. Regul. Integr. Comp. Physiol. 287 (2004) R502–R516.

[7] M.L. Bishop, J.L. Duben-Engelkirk, E.P. Fody (Eds.), Clinical Chemistry: Principles, Procedures, Correlations, second ed., Lippincott, Philadelphia, PA, 1992.

[8] K. Brandis, Acid-base physiology. <http://www.anaesthesiaMCQ.com/AcidBaseBook> (accessed January 2009).

[9] B.A. Shapiro, W.T. Peruzzi, R. Templin, Clinical Applications of Blood Gases, fifth ed., CV Mosby, St. Louis, MO, 1994.

[10] J.G. Toffaletti, Blood Gases and Electrolytes, second ed., AACC Press, Washington, DC, 2009.

[11] H.J. Adrogue, N.E. Madias, Management of life-threatening acid-base disorders, N. Engl. J. Med. 338 (2005) 107–111.

[12] N.A. Mahmood, F.A. Chaudry, H. Azam, M.I. Ali, M.A. Khan, Frequency of hypoxic events in patients on a mechanical ventilator, Int. J. Crit. Illn. Inj. Sci. 3 (2013) 124–129.

[13] K. Wrenn, The delta gap: an approach to mixed acid-base disorders, Ann. Emerg. Med. 19 (11) (1990) 1310–1313.

[14] G. Meisenberg, W.H. Simmons, Principles of Medical Biochemistry, third ed., Saunders (Elsevier), Philadelphia, PA, 2012.

[15] J.M. Robinson, J.R. Lancaster Jr., Hemoglobin-mediated, hypoxia-induced vasodilation via nitric oxide, Am. J. Resp. Cell Mol. Biol. 32 (2005) 257–261.

[16] The ARDS Definition Task Force, Acute respiratory distress syndrome: the Berlin definition, JAMA 307 (23) (2012) 2526–2533.
[17] L. Martin, All You Really Need to Know to Interpret Arterial Blood Gases, second ed., Lippencott Williams & Wilkins, Philadelphia, PA, 1999.
[18] P.D. Stein, S.Z. Goldhaber, J.W. Henry, Alveolar-arterial oxygen gradient in the assessment of acute pulmonary embolism, Chest 107 (1995) 139–143.
[19] A.J. Williams, ABC of oxygen: assessing and interpreting arterial blood gases and acid-base balance, Br. Med. J. 317 (1998) 1213–1216.
[20] J.R. Astles, D. Lubarsky, B. Loun, F.A. Sedor, J.G. Toffaletti, Pneumatic transport exacerbates interference of room air contamination of blood gas samples, Arch. Pathol. Lab. Med. 120 (1996) 642–647.
[21] J.G. Toffaletti, E.H. McDonnell, Effect of small air bubbles on changes in blood pO2 and blood gas parameters: calculated vs measured effects. <http://acutecaretesting.org> (accessed July 2012).
[22] T.J. Mahoney, J.A. Harvey, R.J. Wong, A.L. Van Kessel, Changes in oxygen measurements when whole blood is stored in iced plastic or glass syringes, Clin. Chem. 37 (1991) 1244–1248.
[23] M. Beaulieu, Y. Lapointe, B. Vinet, Stability of pO_2, pCO_2, and pH in fresh blood samples sored in a plastic syringe with low heparing in relation to various blood gas and hematologic parameters, Clin. Biochem. 32 (2) (1999) 101–107.
[24] T.P. Knowles, R.A. Mullin, J.A. Hunter, F.H. Douce, Effects of syringe material, sample storage time, and temperature on blood gases and oxygen saturation in arterialized human blood samples, Respir. Care 51 (7) (2006) 732–736.
[25] B. Fouse, Reference range evaluation for cord blood gas parameters. <www.acutecaretesting.org> (accessed June 2002).
[26] C. Higgins, Umbilical-cord blood gas analysis. <www.acutecaretesting.org> (accessed October 2014).
[27] C. Higgins, Temperature correction of blood gas and pH measurement–an unresolved controversy. <www.acutecaretesting.org> (accessed January 2016).
[28] D. Abramson, T.M. Scalea, R. Hitchcock, S.Z. Trooskin, S.M. Hernry, J. Greenspan, Lactate clearance and survival following injury, J. Trauma. 35 (1993) 584–589.
[29] R.M. Otero, H.B. Nguyen, D.T. Huang, D.F. Gaieski, M. Goyal, K.J. Gunnerson, et al., Early goal-directed therapy in severe sepsis and septic shock revisited: concepts, controversies, and contemporary findings, Chest 130 (2006) 1579–1595.
[30] M. Egi, S. Finfer, R. Bellomo, Glycemic control in the ICU, Chest 140 (2011) 212–220.
[31] NICE Sugar Study Investigators, Intensive versus conventional glucose control in critically ill patients, N. Engl. J. Med. 360 (2009) 1283–1297.
[32] F. Ichinose, J.D. Roberts, W.M. Zapol, Inhaled nitric oxide: a selective pulmonary vasodilator; current uses and therapeutic potential, Circulation. 109 (2004) 3106–3111.

Self-assessment questions

1. Acid–base exercises: In the following table, use the pH, pCO_2, and HCO_3^- values along with the duration of observation to assess the acid−base status of each situation and choose from the answers a−f below. Determine the primary disorder, then evaluate if the expected compensation has occurred for the duration indicated. Use Fig. 36.2 for guidance as needed:

pH	pCO_2 (mmHg)	HCO_3 (mmol/L)	Duration of medical observation	Acid−base condition
7.40	40	24	–	Very normal acid−base results
7.34	32	18	12 h	_____
7.48	45	35	6 h	_____
7.44	27	18	2 d	_____
7.20	70	22	3 d	_____
7.28	55	25	2 h	_____
7.56	20	20	6 h	_____

   a. Acute respiratory acidosis
   b. Acute respiratory alkalosis
   c. Chronic compensated respiratory alkalosis
   d. Mixed respiratory acidosis and metabolic acidosis
   e. Partly compensated metabolic acidosis
   f. Partly compensated metabolic alkalosis
2. Which of the following statements regarding compensation is correct?
   a. Compensation seeks to return the ratio of bicarbonate to H_2CO_3 to normal.
   b. Compensation for respiratory acidosis involves renal excretion of bicarbonate.
   c. Compensation for metabolic alkalosis involves increased respiratory rate.
   d. Compensation typically causes pH to change to the opposite side of normal from that of the original acid−base disorder.
   e. Compensation typically involves altered renal excretion of sodium and potassium.
3. A patient has a COHb of 15%. Which parameter would an increase in COHb affect most?
   a. sO_2
   b. pO_2
   c. pH
   d. $\%O_2Hb$
   e. pCO_2
4. A patient has the following laboratory results (mmol/L): Na 141, K 5.4, Cl 106, and HCO_3 7. The differential for this acid−base abnormality includes all of the following, EXCEPT:
   a. lactate acidosis
   b. diabetic ketoacidosis
   c. severe vomiting
   d. ethylene glycol ingestion
   e. acute renal failure
5. A patient is noted to be hypovolemic, with pH 7.49 and the following electrolyte results (mmol/L): Na 135, K 4.8, Cl 90, and HCO_3 35. What is the most likely diagnosis and appropriate treatment?
   a. Contraction alkalosis; administer intravenous NaCl solution.
   b. Diarrhea; administer HCO_3 solution.
   c. Prolonged vomiting; administer dilute HCl solution.
   d. Respiratory alkalosis; induce hyperventilation.
   e. Respiratory alkalosis; rebreath exhaled air to increase pCO_2.
6. A patient has the following laboratory results: Na 139, K 4.4, Cl 85, HCO_3 42 (all mmol/L), pH 7.35, and pCO_2 79 mmHg. What is the most likely cause of these results?
   a. diabetic ketoacidosis
   b. metabolic alkalosis
   c. acute respiratory acidosis
   d. chronic respiratory acidosis
   e. respiratory alkalosis
7. A patient has an arterial pO_2 of 45 mmHg and a pCO_2 of 80 mmHg. What is the alveolar−arterial gradient (A−a)?
   a. 0 mmHg
   b. 35 mmHg
   c. 10 mmHg
   d. 50 mmHg
   e. 5 mmHg
8. Which condition would most likely increase the pO_2 in an arterial blood sample collected in a syringe from a person breathing room air?
   a. The sample is analyzed 45 minutes after collection and storage at room temperature.
   b. The sample is collected with a pea-sized air bubble in the syringe, then is sent to the lab by pneumatic transport.
   c. The sample is stored on ice for 20 minutes before analysis.
   d. Liquid heparin is used as an anticoagulant.
9. A 60-year-old man has the following lab results:

pH	7.58
pCO_2	35 mmHg
pO_2	95 mmHg
Na	133 mmol/L
K	4.5 mmol/L
Cl	80 mmol/L
HCO_3^-	42 mmol/L.

What is the most likely cause of these results?

a. Compensated metabolic acidosis
b. Five days of severe vomiting
c. Five days of severe diarrhea
d. Acute respiratory alkalosis
e. Chronic respiratory alkalosis

10. Which set of results would most likely be associated with a metabolic acidosis?

a. pH 7.30 pCO_2 55 mmHg HCO_3 27 mmol/L
b. pH 7.40 pCO_2 50 mmHg HCO_3 30 mmol/L
c. Na 136 mmol/L Cl 86 mmol/L HCO_3 36 mmol/L
d. Na 138 mmol/L Cl 87 mmol/L HCO_3 24 mmol/L
e. pH 7.35 pCO_2 30 mmHg HCO_3 17 mmol/L

11. Which cord blood results would NOT likely be a cause for concern?

a. Arterial cord blood pH of 6.96
b. Arterial cord blood pO_2 of 42 mmHg
c. Arterial cord blood pCO_2 of 75 mmHg with a venous cord blood pCO_2 of 40 mmHg
d. Arterial cord blood pH of 7.20 with a venous cord blood pH of 7.40

Answers

1.

pH	pCO_2 (mmHg)	HCO_3 (mmol/L)	Duration of medical observation	Acid–base condition
7.40	40	24	–	Very normal acid–base results
7.34	32	18	12 h	e
7.48	45	35	6 h	f
7.44	27	18	2 d	c
7.20	70	22	3 d	d
7.28	55	25	2 h	a
7.56	20	20	6 h	b

2. a
3. d
4. c
5. a
6. d
7. e
8. b
9. b
10. e
11. b

Chapter 37

Water and electrolyte balance

Anna E. Merrill[1] and Allison B. Chambliss[2]

[1]Department of Pathology, University of Iowa Hospitals and Clinics, Iowa City, IA, United States, [2]Department of Pathology, Keck School of Medicine of USC, Los Angeles, CA, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Describe the distribution of water and major electrolytes among the intracellular and extracellular compartments of the body and the factors involved in their regulation.
- Discuss the clinical evaluation of patients with electrolyte and blood volume disorders and the utility of serum and urine tests in proper diagnosis.
- Explain preanalytical variables that can cause misleading electrolyte results.

Normal physiology of water and electrolytes

Distribution of water and electrolytes between body fluid compartments

In the average adult, water accounts for approximately 60% of the total body weight and exists within three body fluid compartments:

- intracellular fluid (ICF) compartment
- extracellular fluid (ECF) compartments, which consist of plasma and interstitial fluid

The ICF and ECF compartments are separated by cellular plasma membranes, while the capillary endothelium partitions the interstitial fluid from the plasma. The ICF and ECF compartments differ in their water, electrolyte, and protein content. The ICF and ECF compartment compositions are summarized in Table 37.1. Within the ICF, the major cations are potassium (K^+) and magnesium (Mg^{2+}), and the major anion is phosphate (in the forms HPO_4^{2-} and H_2PO_4). The predominant extracellular cations are sodium (Na^+) and calcium (Ca^{2+}), while the predominant extracellular anions are chloride (Cl^-) and bicarbonate (HCO_3^-). This chapter will focus on the analytical and clinical aspects of sodium, potassium, and chloride measurement and regulation. Other electrolytes are discussed elsewhere (see Chapter 36: Blood gas and critical care testing, for additional details on bicarbonate and see Chapter 45: Calcium biology and disorders, for a discussion on magnesium, calcium, and phosphate).

To maintain fluid balance, the average adult must take in ~1.5–2.5 L of water daily to account for renal (1200–1500 mL in urine), fecal (~100 mL in stool), and "insensible" losses (400–700 mL via evaporation from the skin and respiratory tract). Daily water and electrolyte requirements are significantly affected by activity level, environmental conditions, and disease states. Water intake is largely controlled by the sensation of thirst. Ingested fluid first enters the blood plasma by intestinal absorption and then moves between plasma and other compartments based on intercompartmental differences in osmotic, oncotic, and hydrostatic pressures.

Osmotic pressure is caused by a difference in the total molar concentration of compounds dissolved in solutions on either side of a membrane that does not allow for their free passage. Water moves from the compartment with lower total solute concentration to the one with higher total solute concentration to eliminate the imbalance. Osmotic balance is the most important factor regulating movement of water between the plasma and the intracellular spaces, as cell membranes do not permit free diffusion of electrolytes and proteins. The interplay between the hydrostatic and oncotic (colloid osmotic) pressures, along with capillary membrane permeability, determines water distribution between the plasma and interstitial compartments. Hydrostatic pressure exerted by flowing blood drives fluid and diffusible solutes from the vascular space, across the semipermeable capillary membrane, and into the interstitial compartment. Oncotic pressure due to proteins, predominantly albumin, pulls interstitial fluid into the intravascular compartment. Likewise, interstitial hydrostatic and oncotic pressures lead to fluid efflux to and influx from the vascular space, respectively. An imbalance in pressures can favor the accumulation of interstitial fluid (i.e., edema). In addition, certain

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00037-5

TABLE 37.1 Composition of body fluid compartments.

Fluid compartment	Total body water (%)	Major cations	Major anions	Relative protein concentration
Intracellular	50–65	K^+ and Mg^{2+}	$HPO_4^{2-}/H_2PO_4^-$	Very high
Plasma	10–12	Na^+ and Ca^{2+}	Cl^- and HCO_3^-	High
Interstitial	25–40	Na^+ and Ca^{2+}	Cl^- and HCO_3^-	Very low

pathologic conditions (e.g., septic shock) can drastically increase capillary membrane permeability to larger molecules, leading to decreased oncotic pressure, hypotension, and impaired perfusion to multiple organ systems.

The electrolyte compositions of the ECF compartments—plasma and interstitial—are similar, but differ significantly from the ICF compartment. For example, while sodium is the main extracellular cation ($\sim$140 mmol/L in ECF vs. $\sim$12 mmol/L in ICF), potassium is predominantly located intracellularly ($\sim$140 mmol/L in ICF vs. $\sim$4 mmol/L in ECF). Passive ion transport across capillary membranes is responsible for the consistency in electrolyte concentrations between the ECF compartments. The unequal ion distribution between the ICF and ECF compartments is accomplished by active ion transport against concentration gradients. The Na^+/K^+-ATPase enzyme pumps sodium from inside to outside the cell, simultaneously coupling potassium transport from outside to inside the cell.

Regulation of water

To avoid dehydration, daily water intake must replenish losses. Urine represents the main source of water loss, with the kidneys reabsorbing or excreting water based on variation in water intake. The complex regulatory system responsible for maintaining fluid balance relies on various mechanisms for matching water intake with losses and preventing changes in plasma osmolality or volume. In general, preservation of intracellular hydration status is prioritized over extracellular hydration status. Therefore water imbalances are initially reflected in the extracellular component.

Reduced ECF volume can be caused by many conditions, including dehydration, acute blood loss (e.g., trauma), vomiting, diarrhea, diuretics, and renal and adrenal diseases, leading to sodium wasting. Clinically, patients with decreased ECF content demonstrate thirst, anorexia, nausea, lightheadedness, orthostatic hypotension, tachycardia, syncope, oliguria, and decreased skin turgor, with severe cases leading to shock, coma, and death. Regulators of osmolality and volume respond to increases in osmolality and decreases in plasma volume, respectively, to restore ECF fluid balance. Increased ECF volume also has many possible etiologies, including heart failure, cirrhosis, and nephrotic syndrome. In addition, aggressive intravenous (IV) fluid administration can cause iatrogenic fluid overload. Patients with excess ECF typically present with weight gain, edema, dyspnea, tachycardia, jugular venous distention, and ascites due to portal hypertension.

Osmoregulators

The hypothalamus initiates a weak, yet sensitive response to minor increases in plasma osmolality ($\sim$1%) by activating two types of protective mechanisms:

- Thirst sensors stimulate water intake, which lowers osmolality, eventually returning it to normal. Water intake is the most important factor in maintaining normal water and electrolyte status. Thus patients who cannot respond to this signal, such as the elderly, newborns, and those with neurologic disorders or without access to water, are prone to dehydration.
- Antidiuretic hormone (ADH), also known as arginine vasopressin or vasopressin, is released from the posterior pituitary gland and acts on both the kidney and blood vessels. In the kidney, ADH increases the permeability of the collecting ducts to facilitate water reabsorption, thereby lowering plasma osmolality. In the blood vessels, ADH binds to receptors that cause vasoconstriction to maintain blood pressure until fluids are ingested or administered to restore blood volume (see Chapter 39: Disorders of the anterior and posterior pituitary, for further discussion on ADH).

Volume regulators

Normal daily regulation of water status is governed by osmolality, but the body can override the signal of the osmoreceptors, if necessary, to preserve normal plasma volume. Compared with the osmoregulators, the response of the following volume regulators is insensitive (requires a $\sim$5% decrease in plasma volume), yet potent:

- Thirst and ADH release are triggered by a decreased plasma volume, in addition to increases in osmolality.

- The renin–angiotensin–aldosterone system (RAAS) is a major protection against decreased plasma volume. Decreased renal blood flow or sodium concentrations reaching the distal convoluted tubule lead to a release in renin from the juxtaglomerular cells of the kidney. Renin catalyzes production of angiotensin I, which is then converted into angiotensin II. Angiotensin II is a potent vasoconstrictor, which increases renal blood flow, and stimulus for aldosterone secretion from the adrenal cortex. Aldosterone, in turn, acts on the collecting ducts to reabsorb Na^+ and water at the expense of K^+ and H^+ ions. Taken together, the RAAS increases blood pressure, plasma Na^+ retention, and urinary excretion of K^+ and H^+, and decreases urinary Na^+ excretion and plasma concentrations of K^+ and H^+.

Natriuretic peptides

If blood volume becomes abnormally elevated, an increase in cardiac stretch stimulates the production of natriuretic peptides [e.g., atrial natriuretic peptide and brain natriuretic peptide (BNP)] by the atria and ventricles of the heart, respectively. These peptides act on the blood vessels to induce vasodilation and on the kidney to increase urinary excretion of sodium (natriuresis) and water (diuresis). In addition, natriuretic peptides decrease production of renin, serving as an antagonist to the RAAS to decrease plasma volume. Note that plasma levels of BNP or N-terminal pro-BNP are widely used when assessing patients with suspected congestive heart failure, which is a common cause of increased ECF volume and edema (see Chapter 30: Biomarkers for coronary artery disease and heart failure, for further discussion of BNP).

Regulation of electrolytes

Renal regulation, along with thirst, is the body's main mechanism for maintaining water and electrolyte homeostasis. Renal insufficiency limits the ability of the kidneys to regulate the balance of water and electrolytes, which can make affected patients more susceptible to electrolyte abnormalities, particularly hyperkalemia. With normal renal function, the glomeruli freely filter water and electrolytes from the blood. The majority of filtered sodium and potassium are actively reabsorbed in the proximal convoluted tubules; consequently, water and chloride are passively reabsorbed to maintain electrical and osmotic balance. In the descending loop of Henle, an osmotic gradient leads to passive reabsorption of water. In the ascending loop of Henle, chloride and sodium are actively and passively reabsorbed, respectively. The transporters responsible for chloride reabsorption in the ascending loop of Henle include the Na–K–Cl cotransporter—the target of loop diuretics—and CLC chloride channels. Under the action of aldosterone, cells of the distal convoluted tubules and collecting ducts retain Na^+ and water and secrete K^+ and H^+. The collecting duct reabsorbs additional water when stimulated by ADH. In the absence of aldosterone, $\sim$3%–5% of extracellular sodium can be lost in the urine, leading to hyponatremia. In the cases of ADH deficiency, up to 10% of water may be excreted.

Sodium regulation in blood

Sodium is the major extracellular cation, playing a central role in maintaining the normal distribution of water and osmolality in the ECF compartment. Dietary sodium is nearly completely absorbed from the gastrointestinal (GI) tract, though daily intake drastically exceeds requirements in most cases. Excretion of excess sodium is accomplished by the kidneys and is primarily controlled by aldosterone. Plasma sodium concentration is regulated within an extremely narrow window. The average intraindividual variation in healthy people is $<$1.5%. Marked changes in sodium concentration can occur, however, with changes in plasma volume or inability to obtain water.

Potassium regulation in blood

Potassium is the major intracellular cation. Dietary potassium is absorbed in the GI tract, rapidly distributed, and taken up by cells. The kidney principally maintains total body potassium by matching excretion with intake. Rates of renal secretion occur slowly (on the order of hours) and are affected by mineralocorticoid (i.e., aldosterone) concentrations, intake of sodium and potassium, and acid–base balance. More immediate changes in plasma potassium concentrations are buffered by the movement of K^+ between the intracellular and extracellular compartments. Intracellular potassium movement is enhanced by alkalemia, insulin, and glucose. Catecholamines also affect rapid cellular uptake of potassium; catecholamine action on β-adrenergic receptors promotes cellular uptake, while action on α-adrenergic receptors impairs uptake.

Chloride regulation in blood

Chloride is the major extracellular anion; Cl^- ions are almost completely absorbed from the GI tract. Unlike sodium and potassium, concentrations of chloride are not homeostatically controlled; rather, chloride concentrations passively reflect the concentrations of other major ions and typically change in parallel with sodium concentrations. An exception is in acid–base disorders when chloride concentrations shift in response to bicarbonate changes, even if sodium concentrations are normal (e.g., depleted Cl^- due to $HCO3^-$ retention in respiratory acidosis and elevated Cl^- due to $HCO3^-$ excretion in

respiratory alkalosis; see Chapter 36: Blood gas and critical care testing, for further discussion on acid–base disorders).

Laboratory tests used to evaluate fluid and electrolyte abnormalities

When evaluating a patient with fluid and electrolyte abnormalities, the principal laboratory tests include electrolyte concentrations and osmolality, which can be measured in the blood and urine. Though hormones such as ADH, renin, and aldosterone play important roles in maintaining adequate water and electrolyte balance, they are not commonly measured in the context of fluid and electrolyte disorders.

Electrolytes

Electrolytes are typically measured in serum, plasma, whole blood, and/or urine. While expected concentrations of electrolytes are well established for blood-based specimens, urinary concentrations vary widely. However, measuring electrolytes in urine can indicate whether the renal response to changes in plasma concentrations is adequate or whether the kidneys are contributing to an electrolyte imbalance.

Historically, atomic absorption spectroscopy (AAS) and flame emission spectrophotometry (FES) were the most common methods for sodium and potassium analysis. The National Institute of Standards and Technology (NIST), Centers for Disease Control and Prevention (CDC), and American Association for Clinical Chemistry (AACC) recognize FES as the reference measurement procedure for serum sodium and potassium. The most precise method for chloride determination over a wide range of concentrations is coulometric titration, which is the reference measurement procedure for serum chloride designated by the NIST, CDC, and AACC.

In modern clinical laboratories, electrolytes are measured by ion-selective electrode (ISE) methods, which are discussed in detail elsewhere (see Chapter 9: Electrochemistry). These methods utilize potentiometry, the difference in circuit potential induced by the interaction of the selected ion with the ISE membrane, to determine electrolyte concentration. The electrode membranes contain ion-selective compounds, such as glass, valinomycin, or organic polymers. ISE methods are classified into two types: indirect and direct. With indirect ISE methods, the sample is first mixed with a large volume of diluent (typically 1:30–1:40) before introduction into the measurement chamber containing the electrodes. Alternatively, the sample is presented to the electrodes without dilution in direct ISE methods. Most high-throughput clinical chemistry analyzers incorporate indirect ISEs for electrolyte determination in serum, plasma, and urine matrices, while direct ISEs are used more commonly in blood gas and point-of-care devices where whole blood is the specimen of choice for analysis.

Measurements by indirect ISE are vulnerable to inaccuracies due to the electrolyte exclusion effect. Electrolytes that are not protein-bound are confined to the aqueous phase of plasma, which makes up approximately 93% of plasma volume in a normal sample, and are excluded from the solid phase of plasma. Indirect ISE measures the electrolyte concentration in the total plasma volume following a dilution step. To relate the measured concentration to the physiologic electrolyte concentration in plasma water volume, it is assumed that the solid fraction is consistent in volume (7%) (Fig. 37.1, top). In certain pathologic conditions, such as hyperlipidemia or hyperproteinemia, the volume of plasma occupied by solids significantly exceeds 7%. With indirect ISE, this leads to falsely low electrolyte measurements because the total plasma volume is disproportionately excluded by excess solids (e.g., pseudohyponatremia and pseudohypokalemia; see Fig. 37.1, bottom). Conversely, in the settings of decreased solids (e.g., severe hypoproteinemia in critically ill patients), indirect ISE electrolyte measurements will be falsely elevated (e.g., pseudohypernatremia and pseudohyperkalemia). Of note, direct ISE methodologies do not require sample dilution and measure ion concentrations in the aqueous portion of plasma; therefore, they are not affected by deviations in plasma composition. Though the electrolyte exclusion effect affects all electrolytes measured by indirect ISE, pseudohyponatremia and pseudohypernatremia are most noticeable because of the extremely tight regulation of plasma sodium content. For example, a 5% decrease in Na^+ may cause a normonatremic (~138 mmol/L) patient to appear hyponatremic (~131 mmol/L), whereas a 5% decrease in K^+ is clinically less significant (4.0 vs. 3.8 mmol/L).

Despite widespread use of ISE methods due to their robustness and cost-effectiveness, other methods for electrolyte measurement do exist. These methods couple electrolyte-specific enzyme activation with spectrophotometric detection. Enzymatic/spectrophotometric techniques are typically incorporated into smaller instruments with Clinical Laboratory Improvement Amendments (CLIA)-waived status for whole blood specimens, making them attractive for environments with limited testing resources. A few settings where these methods may be incorporated include physicians' offices and isolation laboratories for patients with highly infectious diseases (e.g., Ebola).

Sodium

A typical reference interval for serum or plasma sodium is 135–145 mmol/L (due to its univalency, 1 mmol/L

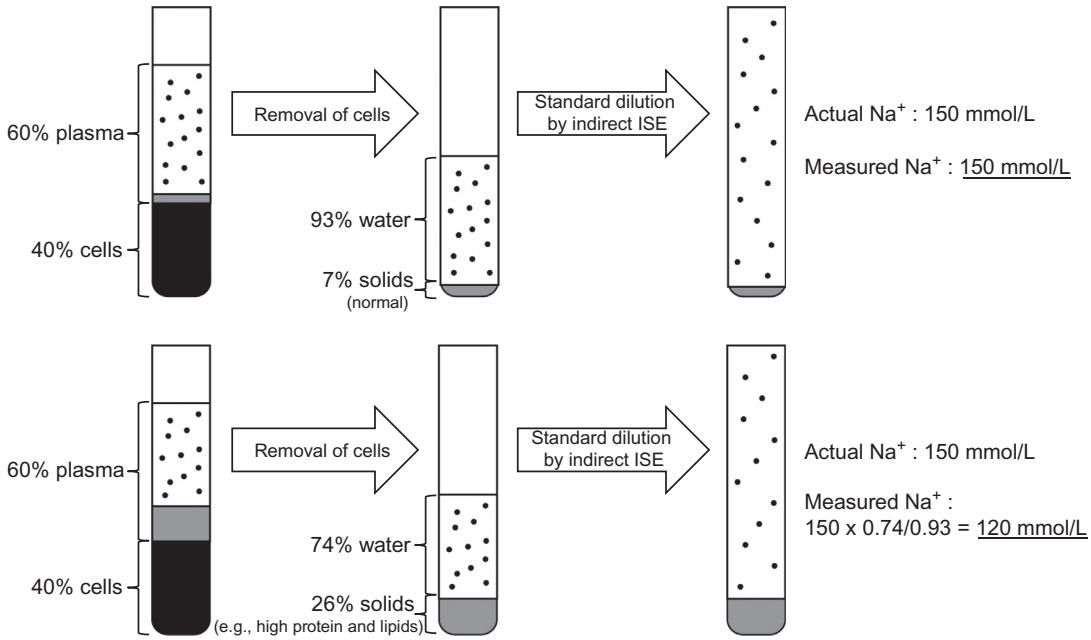


FIGURE 37.1 Example of dilution scheme for the indirect ion-selective electrode measurement of sodium in two samples with the same sodium concentration in plasma water but with different proportions of plasma solids: normal (7% solids, *top*) and elevated (26% solids, *bottom*). Although the two samples started with the same sodium concentration, the sodium result of the sample with elevated solids is significantly underestimated by the indirect ion-selective electrode method due to the electrolyte exclusion effect, demonstrating pseudohyponatremia.

$Na^+ = 1$ mEq/L Na^+). Urinary excretion of sodium varies with dietary intake, but a typical observed range is approximately 120−240 mmol/day. Determination of the fractional excretion of sodium may be more useful than urine sodium concentration alone in patients with acute kidney injury to differentiate prerenal disease and acute tubular necrosis as possible etiologies. This calculation provides an index of the urine sodium (U_{Na}) relative to the plasma sodium (P_{Na}) in proportion to the ratio of plasma (P_{Cr}) to urine creatinine (U_{Cr}). The formula for fractional excretion of sodium (FE_{Na}), expressed as a percent, is given in Eq. (37.1):

$$FE_{Na} = \frac{U_{Na} \times P_{Cr}}{P_{Na} \times U_{Cr}} \times 100 \qquad (37.1)$$

FE_{Na} of $<1\%$ typically indicates prerenal disease, while FE_{Na} of $>2\%$ indicates acute tubular necrosis; however, these classifications are of limited specificity, especially in patients taking diuretics or with other causes of sodium wasting.

Potassium

A typical reference interval for serum or plasma potassium is approximately 3.5−5.1 mmol/L (1 mmol/L $K^+ = 1$ mEq/L K^+). Urinary excretion of potassium varies with dietary intake, but a typical observed range is approximately 20−80 mmol/day. Serum and plasma potassium is particularly prone to collection-related pseudohyperkalemia, in large part, because intracellular K^+ concentrations in RBCs, white blood cells, and platelets exceed plasma K^+ concentrations by over 30-fold. Thus even minor cell lysis during the preanalytical phase can lead to falsely elevated potassium concentrations in the plasma. The most common cause of collection-related pseudohyperkalemia is hemolysis. Pseudohyperkalemia can also arise from the contamination of the specimen with potassium-containing IV fluids or an improper anticoagulant (e.g., K_2- or K_3-EDTA). Various sources of pseudohyperkalemia, along with mechanisms of interference and tips for prevention, are listed in Table 37.2.

Chloride

A typical reference interval for serum or plasma chloride is approximately 98−111 mmol/L (1 mmol/L $Cl^- = 1$ mEq/L Cl^-). Urinary excretion of chloride varies with dietary intake, but a typical observed range is approximately 110−250 mmol/day. Chloride ion concentrations are also determined in sweat specimens in the diagnosis of cystic fibrosis (CF), usually following a positive CF newborn screening result. Sweat production is stimulated by pilocarpine iontophoresis, and the collected specimens are submitted to the laboratory for quantitative chloride analysis. Sweat chloride concentrations of 60 mmol/L or higher are indicative of CF, while concentrations less than 30 mmol/L indicate that CF is unlikely. Given the expected chloride concentrations in sweat specimens from

TABLE 37.2 Causes of pseudohyperkalemia and tips for their prevention.

Cause	Mechanism	Tips for prevention
Traumatic phlebotomy	Release of intracellular K^+ from prolonged tourniquet application and/or fist clenching	Practice proper phlebotomy technique
Hemolysis	Release of K^+ from red cells due to mechanical lysis	Collect blood via venipuncture instead of indwelling lines Use larger bore/smaller gauge needles Hand-carry samples to laboratory
Improper order of draw or tube type	Contamination with K_2- or K_3-EDTA anticoagulant	Adhere to proper order of draw and tube type
Contamination with IV fluid	Contamination with K^+ replacement infusion	Collect blood via venipuncture Properly flush IV with multiple blood volumes prior to specimen collection
Delayed separation of serum/plasma from cellular material	Leakage of K^+ from blood cells due to time- and temperature-dependent inactivation of Na^+/K^+-ATPase	Promptly aliquot serum/plasma after collection Utilize serum or plasma separator tubes (SST or PST) and promptly centrifuge after collection
Thrombocytosis	Release of K^+ from platelets during clotting	Measure potassium in plasma specimen
Leukocytosis	Release of K^+ from white blood cells due to mechanical disruption and/or extremely fragile cells	Measure potassium in serum or whole blood specimen Hand-carry samples to laboratory

IV, Intravenous.

this patient population, it is recommended that instrumentation used for measuring Cl^- in sweat be capable of reliably measuring concentrations from 10 to 160 mmol/L. ISE methods are usually too imprecise at the low end of this concentration range. Therefore sweat chloride concentration is best determined by coulometric titration on specialized instruments.

Osmolality

Determination of serum/plasma and urine osmolality can be useful in the assessment of fluid and electrolyte disorders. Specifically, measured osmolalities in these two specimen sources can reveal the status of renal water regulation in settings of severe electrolyte disturbances [e.g., diabetes insipidus (DI) and syndrome of inappropriate antidiuretic hormone secretion (SIADH)]. In addition, the difference between the measured osmolality and the estimated osmolality, typically calculated using sodium, glucose, and blood urea nitrogen (BUN) concentrations (see Chapter 6: Laboratory calculations, for a discussion of this calculation), is known as the osmolality gap. An elevated osmolality gap suggests the presence of additional osmotically active substances, such as methanol, ethylene glycol, or ketoacids. A thorough clinical history and/or more specific testing are often needed to identify the source of the elevated osmolality gap. Osmolality can theoretically be measured by any of the four colligative properties of solutions: increased osmotic pressure, decreased vapor pressure, increased boiling point, and decreased freezing point. The magnitude of change in any of these parameters relative to water is directly proportional to the total number of solute particles dissolved in the solvent. In the clinical laboratory, osmolality is most commonly measured by freezing point depression. A typical reference interval for serum or plasma osmolality is approximately 275−300 mOsm/kg. In urine, expected osmolality is variable due to diet and hydration state but is often approximately 350−1050 mOsm/kg. Normally, the ratio of urine osmolality to plasma osmolality is typically 1.0−3.0.

Disorders of water and sodium

Abnormalities in plasma or serum sodium concentrations may be caused by disorders of sodium regulation or by disorders of body water regulation, producing a dilutional effect in sodium. Clinical signs indicating possible disorders of water or sodium include, for example, the presence of generalized or localized edema and/or low blood

pressure in hypovolemic patients. Nonspecific symptoms of hyponatremia may include weakness and fatigue, and, in severe cases, confusion or coma. However, slow changes in sodium concentrations may not be apparent; thus sodium is often monitored as part of a metabolic panel regardless of the presence of symptoms. Patients with hypernatremia may exhibit thirst and dry mucous membranes.

Obtaining a clinical history including recent diet, fluid intake, history of vomiting or diarrhea, kidney function status, and use of medications such as diuretics is critical to the workup of patients exhibiting electrolyte imbalance. Measurement of BUN and creatinine may aid in evaluating intravascular volume and determining renal function. An increased BUN/creatinine ratio may indicate decreased intravascular volume in cases of hyponatremia. Further, measurement of serum and urine osmolality may aid in determining the etiology of a sodium abnormality. Caution should be exhibited in the interpretation of sodium results from indirect ISE methods in patients with extreme plasma protein and/or lipid concentrations, as the plasma water displacement may cause pseudohypernatremia (with decreased solids) or pseudohyponatremia (with increased solids), as discussed above.

Disorders causing hyponatremia

Hyponatremia has been described as the most common electrolyte disorder encountered in clinical practice, occurring in an estimated 15%–30% of hospitalized patients. While many cases are mild and asymptomatic, occurrences that progress unrecognized and untreated may cause significant morbidity and mortality. Acute hyponatremia due to osmotically induced brain edema represents one of the most severe manifestations. Hyponatremia is typically defined as a serum or plasma sodium concentration less than 135 mmol/L. Causes of hyponatremia may be broadly categorized as disorders of decreased ECF sodium or disorders of increased ECF water. Various etiologies for both categories are described below and summarized in Table 37.3. Notably, healthy individuals possess compensatory mechanisms, such as thirst and ADH, to preserve normal sodium and body water.

Many cases of decreased plasma sodium are considered hypovolemic hyponatremia, by which the hyponatremia is simultaneously associated with a decrease in intravascular volume, and the sodium loss is in excess relative to the water loss. Hypovolemic hyponatremia may be the result of decreased sodium intake, extrarenal sodium losses, renal sodium losses, or increased ECF water. However, some cases of hyponatremia may present with no relative change in intravascular volume (euvolemic) or, rarely, with increases in intravascular volume (hypervolemic). Euvolemic or hypervolemic hyponatremia occurs when effective solutes other than sodium overpower plasma osmolality, causing osmotic movement of water out of cells and a dilutional decrease in plasma sodium. Hypervolemic hyponatremia is most commonly caused by severe hyperglycemia in diabetic patients, though similar effects have been observed after the administration of hypertonic mannitol or glycine.

Decreased sodium

Rarely, decreased intake of dietary sodium will result in hyponatremia. This cause may be considered in alcoholics or in those with anorexia nervosa. Extrarenal, or nonrenal,

TABLE 37.3 Summary of common causes of hyponatremia with typical laboratory features.

Etiology	BUN/Cr ratio	Plasma osmolality	Urine osmolality	Other notes
Osmotic dilution	↑	↑	Varies	Most commonly occurs with hyperglycemia; may rarely occur from mannitol or glycine
Nonrenal sodium loss	↑	↓	↑	GI (diarrhea) or skin (burns and sweat)
Renal sodium loss	↑	↓	↑	Common with thiazide diuretics; may rarely occur with loop diuretics or tubular damage/renal disease
Decreased aldosterone	↑	↓	↑	Adrenal insufficiency or drugs, which inhibit renin or aldosterone, and hyperkalemia and nonanion gap metabolic acidosis also present
Water overload	↑	↓	↓	Psychogenic polydipsia, over-replacement after strenuous exercise
SIADH	↓	↓	↑	Typically euvolemic by physical examination

ADH, Antidiuretic hormone; *SIADH*, syndrome of inappropriate antidiuretic hormone.

losses of sodium may occur through the GI tract (diarrhea) or through the skin (burns and sweat). These patients typically show relatively low excretion of sodium in the urine (spot urine Na^+ often <30 mmol/L) with otherwise high urine osmolality, often >500 mOsm/kg.

Significant sodium loss through the kidneys is most commonly due to the administration of diuretic therapies, such as those used in the treatment of hypertension. In particular, thiazide diuretics inhibit the Na^+/Cl^- cotransporter in the distal convoluted tubule and ultimately decrease sodium reabsorption. Risk factors such as old age, low body weight, and concurrent medications that impair water excretion may further contribute to thiazide-induced hyponatremia. In contrast, hyponatremia is rarely induced by loop diuretics, which impair Na^+/Cl^- transport in the ascending loop of Henle. At this location of the nephron, water retention is limited. A less common cause of renal sodium loss is intrinsic renal disease due to tubular damage or interstitial nephritis. Patients with renal sodium losses tend to show relatively elevated urinary sodium excretion (spot urine Na^+ often >30 mmol/L) with high urine osmolality.

Rarely, a deficiency of renin or aldosterone is the cause of renal sodium loss. In normal states, renin stimulates aldosterone secretion, and aldosterone subsequently increases sodium reabsorption via the expression of epithelial sodium channels in the distal tubule. Renin may be decreased in individuals with intrinsic renal damage caused by diabetes or interstitial nephritis, or in those taking certain medications such as nonsteroidal anti-inflammatory agents (NSAIDs) or calcineurin inhibitors. Aldosterone deficiency may occur with medications that inhibit its production, including angiotensin inhibitors and heparin. Hyponatremia from hypoaldosteronism may also be caused by primary adrenal insufficiency. Patients with significant deficiencies in renin and aldosterone exhibit high sodium excretion in urine, hyperkalemia, and nonanion gap metabolic acidosis.

Increased water

In addition to osmotic shifts caused by hyperglycemia (or rarely, mannitol or glycine), hyponatremia associated with increased water (dilutional hyponatremia) may be the result of excess water ingestion or excess secretion of ADH. Increased water retention often manifests clinically as edema. Excess water ingestion is a rare cause of hyponatremia but may occur if patients drink large amounts of water in a short amount of time. Excess water ingestion has been observed in psychiatric disorders (psychogenic polydipsia) or in patients attempting to replace fluid losses after strenuous exercise. These patients present with water-like urine with extremely low urine osmolality, electrolytes, and creatinine.

In a normal state, ADH production at the hypothalamus and secretion at the posterior pituitary gland are stimulated by increased plasma osmolality and by decreased blood volume. Once secreted, ADH acts at the collecting ducts of the kidneys to facilitate antidiuresis in order to restore plasma osmolality and blood volume. SIADH involves an uncompensated release of ADH despite decreased osmolality and normal to increased intravascular volume, features that should normally inhibit ADH production. The excess ADH stimulates continuous renal retention of water, and the patient will exhibit decreased plasma osmolality with urine that is inappropriately concentrated. Typical findings include urine osmolality >200 mOsm/kg without evidence of volume depletion. Measurement of plasma ADH itself is not often routinely available and is typically not necessary for diagnosis. Further, ADH concentrations are variably elevated in patients with SIADH, are often elevated in all types of hyponatremia, and are affected by many types of medications. Causes of excess ADH with accompanying hyponatremia are listed in Table 37.4 and include ectopic production, drugs such as carbamazepine, and adrenal insufficiency.

TABLE 37.4 Example conditions leading to excess or deficient ADH release and possible hypo- or hypernatremia.

Excess ADH → hyponatremia	ADH deficiency or resistance → hypernatremia

• Ectopic production in tumors (SIADH), for example, small cell lung cancer
• Drug-induced, for example, carbamazepine, diuretics
• Adrenal insufficiency (reduces blood pressure and cardiac output)
• Hypothyroidism (inhibits free water excretion)
• Alcohol withdrawal
	• Central DI
	  ◦ Pituitary tumors
	  ◦ Neurosurgery, stroke, or head trauma
	• Nephrogenic DI
	  ◦ Chronic lithium therapy
	  ◦ Persistent hypercalcemia
	• Granulomatous disease
	• Idiopathic, possibly autoimmune
	• Familial, autosomal dominant (rare)

ADH, Antidiuretic hormone; *DI*, diabetes insipidus; *SIADH*, syndrome of inappropriate antidiuretic hormone.

Importantly, in most cases of hyponatremia, sodium administration is not the mainstay of treatment. In patients with osmotic dilutions, such as in the case of hyperglycemia, plasma sodium will increase as the other solute resolves. In cases of excess water ingestion, fluid restriction restores the plasma sodium imbalance. Similarly, water restriction is the most common approach to SIADH, as administration of sodium will serve only to raise the urine sodium without correction of the plasma sodium in these patients. Hypertonic saline administration is generally considered only in severe, symptomatic cases of hyponatremia. The saline must be administered slowly and cautiously to avoid osmotic demyelination syndrome (ODS). Various sodium administration protocols are established depending on the patient's risk for ODS, which increases with severe hyponatremia (plasma sodium ≤ 105 mmol/L), hypokalemia, alcoholism, malnutrition, and advanced liver disease.

Disorders causing hypernatremia

Hypernatremia, typically defined as a serum or plasma sodium concentration greater than 145 mmol/L, occurs less frequently than hyponatremia. Because hypernatremia is avoided with adequate water intake, hypernatremia most often occurs in those without free access to water, such as infants or bedridden patients, or in those whose thirst mechanism does not function properly, such as the elderly and those with neurologic disorders. Patients with hypovolemic hypernatremia exhibit signs of dehydration with concentrated urine and increased BUN. A low urine sodium indicates extrarenal water losses and an appropriate response by the kidneys to conserve sodium and water in this setting. Treatment seeks to replace free water with administration of dilute oral or IV fluids. Once intravascular volume has been restored, BUN normalizes and sodium will fall.

Less commonly, hypernatremia may occur in the setting of normal ECF volume in patients with DI, a condition defined by inadequate production or action of ADH. DI is classified as either central or nephrogenic. Central DI is an ADH deficiency caused by CNS abnormalities that affect ADH production at the hypothalamus, such as tumors, stroke or head trauma, or granulomatous diseases. Causes of ADH deficiency with accompanying hypernatremia are summarized in Table 37.4. Nephrogenic DI describes a renal tubular resistance to ADH, which may be hereditary if present in children, such as from mutations in the vasopressin V2 receptor gene, *AVPR2*. In adults, nephrogenic DI may be caused by some drugs, such as in chronic lithium therapy, or persistent hypercalcemia. Despite the cause, DI manifests as polyuria with an extremely dilute urine (osmolality typically <100 mOsm/kg). In patients with an adequate thirst mechanism, fluid intake will offset much of the free water loss and hypernatremia will remain mild. Thus patients may only exhibit symptoms of polyuria and polydipsia. However, without access to water, patients will quickly become dehydrated, and overt hypernatremia may develop. Although not typically necessary for diagnosis, a water deprivation test may be performed, which will show no responsive increase in urine osmolality in patients with DI. Further, administration of ADH serves to differentiate central from nephrogenic DI; only patients with central DI will respond appropriately to the exogeneous ADH.

Rarely, hypervolemic hypernatremia may result from the administration of hypertonic fluids, such as hypertonic sodium or sodium bicarbonate, or in cases of accidental or surreptitious salt poisoning of infants or young children. In the worst cases, a rapid hypernatremia will develop with ODS. In these patients, free water replacement is imperative, or possibly dialysis in patients with acute kidney injury.

Case example

A 43-year-old male patient with focal epilepsy, who recently started a new therapy regimen due to increased frequency of seizures, visited his general practitioner for recurrent episodes of nausea and dizziness. A review of recent clinical and lab history suggested normal renal function with no thyroid or adrenal disorders. Labs at this visit indicated hyponatremia (plasma sodium of 128 mmol/L by indirect ISE) with no change when repeated with a direct ISE method. Upon further workup, urine sodium resulted as 45 mmol/L with a urine osmolality of 212 mOsm/kg. No abnormalities in kidney function or acid–base balance were observed. What is the likely etiology of the hyponatremia?

Case resolution

In light of recent medication changes, a review of medication history is warranted in the workup of new onset hyponatremia. This patient was recently prescribed carbamazepine for increased frequency of seizures. Carbamazepine, which is known to induce ADH release, likely caused excessive water retention and hyponatremia in this patient. Resolution of sodium concentrations upon cessation of carbamazepine would confirm the diagnosis.

Disorders of potassium

Unlike sodium, potassium abnormalities typically occur without changes in body water or volume. As potassium is the major intracellular cation, potassium abnormalities may result from disruptions in the balance of its intake, excretion, or shifts into and out of cells.

Disorders causing hypokalemia

Symptoms of hypokalemia often do not manifest until plasma potassium concentrations fall below 3.0 mmol/L and may include muscle weakness and irritability. Below 2.5 mmol/L, cardiac arrhythmias and electrocardiogram abnormalities are observed, such as prolonged QT interval. The evaluation of a patient with hypokalemia should consist of a review of history for diarrhea, vomiting, and use of diuretics and may necessitate assessment of urinary potassium excretion along with acid–base status. Hypokalemia is most often caused by renal potassium losses but may also result from significant shift of potassium out of the vasculature and into cells. These two causes are discussed in more detail below. Other causes of hypokalemia may include decreased dietary intake in malnutrition and GI losses from diarrhea and vomiting or nasogastric suction. Oral or IV potassium replacement is the mainstay of treatment for hypokalemia, in addition to treating the underlying cause. Various causes and mechanisms of hypokalemia are summarized in Table 37.5.

Renal potassium loss

Most commonly, loss of potassium through the kidneys is the cause of hypokalemia. Increased mineralocorticoid or glucocorticoid activity, such as in cases of hyperaldosteronism or Cushing's syndrome, will enhance aldosterone-mediated sodium reabsorption at the expense of potassium. These patients will show significant potassium in the urine in addition to their hypokalemia. Hypomagnesemia is often associated with hypokalemia and may itself cause hypokalemia; depletion of magnesium counteracts magnesium's inhibitory effect on luminal potassium channels. Another frequent cause of renal potassium loss is the use of loop and thiazide diuretics, which activate the RAAS and increase potassium excretion. If potassium intake is not sufficient to compensate, hypokalemia will occur. Bartter and Gitelman syndromes are inherited disorders that mimic the effects of loop and thiazide diuretics, respectively, and may cause hypokalemia. In addition, a common side effect of the antifungal medication amphotericin B is increased tubular membrane permeability to potassium, also causing hypokalemia.

TABLE 37.5 Example causes and mechanisms of hypokalemia.

Reduced potassium intake
• Malnutrition and poor diet (chronic)
Nonrenal potassium losses
• Diarrhea, vomiting, or nasogastric suction
Renal potassium losses
• Hyperaldosteronism or hypercortisolism
• Magnesium deficiency
• Loop and thiazide diuretics
• Bartter and Gitelman syndromes
• Amphotericin B
Increased cellular uptake
• Alkalemia
• Administration of insulin or glucose
• Excess of catecholamines or β adrenergic agonists

Shift of potassium into cells

Normal balance of potassium between cells and the ECF is maintained primarily by Na^+/K^+-ATPases within the cell membrane, which exchange intracellular Na^+ for extracellular K^+. Increased activation of this exchange causes increased potassium shift into cells that will decrease potassium in the plasma. Two major activators of such a shift toward intracellular potassium include alkalemia and the administration of insulin or glucose. In alkalemia, activity of the Na^+/K^+-ATPases is promoted and shifts plasma potassium into cells. Alkalemia-mediated hypokalemia is transient and will resolve when the underlying acid–base disorder is treated. Notably, the reverse mechanism may also occur; hypokalemia from other causes may lead to metabolic alkalosis. Insulin also activates the Na^+/K^+-ATPases. Patients presenting with diabetic ketoacidosis who are administered exogenous insulin often exhibit a significant, but transient, decline in serum/plasma potassium. In addition, elevated β-adrenergic activity via endogenous catecholamines is also known to shift potassium into cells.

Disorders causing hyperkalemia

Hyperkalemia may be physiological or artifactual, owing to preanalytical specimen issues. Like hypokalemia, clinical hyperkalemia often manifests with muscle weakness and paralysis. Prolonged concentrations above 7.0 mmol/L typically cause cardiac conduction abnormalities, such as peaked T waves on electrocardiogram. Hyperkalemia in this range, if true, is potentially fatal and requires emergent treatment. Etiologies of hyperkalemia include increased potassium intake, increased retention (decreased renal excretion), or redistribution of potassium out of cells. Increased intake alone, either by diet or supplements, does not typically cause hyperkalemia in the setting of normal renal function and potassium excretion. The evaluation of a patient with hyperkalemia should include a careful review of medication history and renal

TABLE 37.6 Example causes and mechanisms of hyperkalemia.

Increased potassium load (must have diminished renal function) • Intravenous (KCl or transfusions) • Oral intake (dietary supplements)
Decreased renal excretion • Acute and chronic kidney disease • Adrenal deficiency, Addison's disease • Drug-induced: ACE inhibitors, β-adrenergic blockers, and NSAIDs • Potassium-sparing diuretics (e.g., spironolactone)
Cellular loss • Acidemia: inhibits Na^+/K^+-ATPases • Digoxin toxicity: also inhibits Na^+/K^+-ATPases • Rapid cell destruction: rhabdomyolysis, hemolytic anemia, and tumor lysis syndrome • Insulin deficiency: blocks entry of K^+ into cells
Preanalytical, collection-related, and artefactual • Hemolysis, cell lysis, delayed centrifugation, and K_2- or K_3-EDTA contamination

ACE, Angiotensin-converting enzyme; *NSAID*, nonsteroidal antiinflammatory agents.

function. Various causes and mechanisms of hyperkalemia are summarized in Table 37.6.

Notably, preanalytical factors contributing to pseudohyperkalemia should be considered upon findings of unexpectedly elevated potassium results (Table 37.2). Failure to identify high potassium results as artifactual may result in unnecessary treatment that leads to dangerous hypokalemia; conversely, true hypokalemia that is viewed as within normal limits by preanalytic effects may prevent patients from receiving necessary potassium supplementation.

Decreased renal excretion

Decreased renal excretion is a common cause of hyperkalemia and may be exacerbated by drugs that have effects of reduced renal potassium clearance. Hyperkalemia is a frequent complication of acute and chronic kidney disease. Adrenal deficiency, such as in the setting of Addison's disease, will diminish aldosterone-mediated sodium reabsorption, and thus results in increased potassium retention at the tubules. Further, drugs that inhibit the RAAS, including angiotensin-converting enzyme (ACE) inhibitors, β-adrenergic blockers, and NSAIDs, also commonly cause hyperkalemia and should be discontinued if possible in the setting of severe hyperkalemia. Potassium-sparing diuretics (e.g., spironolactone) administered in excess may also cause hyperkalemia.

Shift of potassium out of cells

States of acidemia inhibit the Na^+/K^+-ATPase exchange, will shift potassium extracellularly, and may cause hyperkalemia in the absence of urinary or GI losses. As with alkalemia-mediated hypokalemia, acidemia-mediated hyperkalemia is transient and will resolve when the underlying acid–base disorder is treated. Cardiac glycosides such as digoxin function by inhibiting Na^+/K^+-ATPase activity; thus hyperkalemia is observed in cases of digoxin toxicity. Conditions that cause rapid cell destruction, such as rhabdomyolysis, hemolytic anemia, and tumor lysis syndrome, will also cause significant potassium release into the plasma, though hyperkalemia may not be observed unless renal function is insufficient. Finally, insulin deficiency prevents potassium from entering cells and contributes to increased plasma potassium concentrations. Patients with an acute hyperkalemic presentation may be administered insulin to facilitate rapid intracellular transfer of potassium.

Case example

A 76-year-old female was visited by a nurse as part of a home health care service for routine diabetes management. Otherwise, the patient appeared well, and blood was drawn for a basic metabolic panel. Later that evening, the patient's physician called the patient and instructed her to go to her closest emergency department (ED) due to an alarmingly high plasma potassium result of 6.3 mmol/L. The physician also noted moderately low glucose. Upon retesting in the ED, the patient's potassium resulted as 4.1 mmol/L, and she was discharged home. What was the cause of the discrepant potassium results?

Case resolution

After visiting the patient, the home health nurse visited several other patients over several hours prior to dropping off blood specimens in batch at the laboratory. The blood tube was not centrifuged in the meantime, allowing intracellular potassium to leak out into the plasma due to progressively reduced activity of Na^+/K^+-ATPases as glucose was depleted by ongoing glycolysis. The potassium result from the sample that was processed immediately after draw at the ED lab should be considered the preferred (true) result.

Disorders of chloride

Unlike sodium and potassium, abnormalities of plasma chloride concentrations alone are not associated with severe clinical consequences. However, assessment of chloride is useful in the differential diagnosis of acid–base disorders and in the determination of the anion gap. In the absence of acid–base disorders, chloride concentrations tend to follow the trends of sodium. In addition, CF is associated with mutations in the CF transmembrane regulator gene, which codes for chloride channels across epithelial membranes. Elevated chloride in sweat remains a mainstay diagnostic indicator for CF.

Disorders causing hypochloremia

Generally, disorders causing hyponatremia may cause concurrent hypochloremia. In particular, prolonged vomiting may deplete chloride to a higher degree than sodium due to the high proportion of chloride in gastric fluid. Hypochloremia may result from renal losses that also deplete sodium, such as osmotic dilution or use of diuretics. Hypochloremia with normal plasma sodium is observed in respiratory alkalosis, which may occur, for example, in the cases of pulmonary edema, chronic obstructive pulmonary disease, or asthma. Acute exacerbations of these conditions cause increased secretion of hydrogen ion at the proximal tubules, and sodium is retained preferentially as sodium bicarbonate rather than sodium chloride. Administration of chloride-containing solutions may be necessary in the cases of true chloride depletion.

Disorders causing hyperchloremia

Causes of hypernatremia may also cause hyperchloremia, such as dehydration or DI in the absence of water replacement. Overuse of normal saline solution, which has a chloride concentration of 154 mmol/L, may cause a hyperchloremia metabolic acidosis as bicarbonate is increasingly excreted. By the same mechanism, plasma chloride may rise in cases of compensated respiratory alkalosis.

Acknowledgments

The authors gratefully acknowledge the authors of the previous edition of the Water and electrolytes balance chapter, John Toffaletti and D. Robert Dufour, on which parts of this chapter are based.

Suggested reading

[1] J. Toffaletti, D.R. Dufour, Water and electrolyte balances, in: W. Clarke (Ed.), Contemporary Practice in Clinical Chemistry, third ed., AACC Press, Washington, DC, 2016, pp. 467–476.

[2] E.I. Schindler, S.M. Brown, M.G. Scott, Electrolytes and blood gases, in: N. Rifai, A.R. Horvath, C.T. Wittwer (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, sixth ed., Saunders Elsevier, St. Louis, MO, 2017, pp. 604–625.

[3] Higgins, C. Pseudohyponatremia. <http://acutecaretesting.org>, (accessed January 2019).

[4] Q.H. Meng, E.A. Wagar, Pseudohyperkalemia: a new twist on an old phenomenon, Crit. Rev. Clin. Lab. Sci. 52 (2) (2015) 45–55.

[5] S. Cheng, E.I. Schindler, M.G. Scott, Disorders of water, electrolytes, and acid–base metabolism, in: N. Rifai, A.R. Horvath, C.T. Wittwer (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, sixth ed., Saunders Elsevier, St. Louis, MO, 2017, pp. 1324–1347.

[6] M. Reynolds Rebecca, L. Padfield Paul, R. Seckl Jonathan, Disorders of sodium balance, BMJ 332 (2006) 702.

[7] J.G. Verbalis, S.R. Goldsmith, A. Greenberg, et al., Diagnosis, evaluation, and treatment of hyponatremia: expert panel recommendations, Am. J. Med. 126 (Suppl. 1) (2013) S1–S42.

[8] K.S. Hwang, G.H. Kim, Thiazide-Induced Hyponatremia, Electrolyte Blood Press 8 (1) (2010) 51–57.

[9] Clinical Practice Guidelines: Treatment of Acute Hyperkalemia in Adults. UK Renal Association, March 2014.

[10] G. Morrison, Serum chloride, in: H.K. Walker, W.D. Hall, J.W. Hurst (Eds.), Clinical Methods: The History, Physical, and Laboratory Examinations, third ed., Butterworths, Boston, MA, 1990.

[11] P.M. Farrell, T.B. White, C.L. Ren, et al., Diagnosis of cystic fibrosis: consensus guidelines from the Cystic Fibrosis Foundation, J. Pediatric. 181S (2017) S4–S15.

Self-assessment questions

1. The renin–angiotensin–aldosterone system serves which of the following functions?
   - **a.** plasma K^+ retention, urinary excretion of Na^+, and increase blood pressure
   - **b.** plasma K^+ retention, urinary excretion of Na^+, and decrease blood pressure
   - **c.** plasma Na^+ retention, urinary excretion of K^+, and increase blood pressure
   - **d.** plasma Na^+ retention, urinary excretion of K^+, and decrease blood pressure
   - **e.** plasma retention and urinary excretion of both Na^+ and K^+, increase blood pressure
2. Which of the following techniques is a modern method for measurement of electrolytes that is preferred in cases of increased blood solids (e.g., hyperproteinemia and hyperlipidemia)?
   - **a.** direct ISE potentiometry
   - **b.** indirect ISE potentiometry
   - **c.** FES
   - **d.** AAS
   - **e.** coulometric titration
3. Which of the following techniques is preferred for the measurement of sweat chloride due to its precision at low chloride concentrations?
   - **a.** direct ISE
   - **b.** indirect ISE
   - **c.** FES
   - **d.** AAS
   - **e.** coulometric titration
4. Of the following causes of hyponatremia, which would most likely present with a high serum osmolality?
   - **a.** renal fluid and electrolyte losses
   - **b.** psychogenic polydipsia
   - **c.** syndrome of inappropriate antidiuresis (SIADH)
   - **d.** edematous states such as congestive heart failure
   - **e.** diabetes mellitus with extreme hyperglycemia
5. Which of the following conditions is likely to cause hypernatremia due to a deficiency of ADH (central diabetes insipidus)?
   - **a.** syndrome of inappropriate antidiuresis (SIADH)
   - **b.** use of diuretics
   - **c.** adrenal insufficiency
   - **d.** pituitary tumor
   - **e.** hypothyroidism
6. A patient with normal potassium 3 months ago is found to have a serum K^+ of 3.1 mmol/L. He had recently been started on several medications. From the following list, which medication is the most likely to have contributed to hypokalemia in this patient?
   - **a.** ibuprofen (NSAID)
   - **b.** enalapril (ACE)
   - **c.** furosemide (loop diuretic agent)
   - **d.** metoprolol (B-adrenergic blocking agent)
   - **e.** spironolactone (aldosterone antagonist)
7. Which of the following preanalytical factors contribute to falsely elevated potassium results?
   - **a.** delayed or incomplete separation of serum from cells
   - **b.** contamination with K_2- or K_3-EDTA anticoagulant
   - **c.** prolonged fist clenching during blood collection
   - **d.** a and b
   - **e.** all of the above
8. Blood chloride concentrations tend to follow trends of _____ but may deviate in certain acid–base disorders.
   - **a.** sodium
   - **b.** potassium
   - **c.** hydrogen
   - **d.** bicarbonate
   - **e.** magnesium

Answers

1. c
2. a
3. e
4. e
5. d
6. c
7. e
8. a

Chapter 38

Urinalysis

Shu-Ling Fan[1,2] **and Shi Bai**[1,2]

[1]*Department of Pathology, UMass Memorial Medical School, Worcester, MA, United States,* [2]*Department of Pathology, UMass Memorial Medical Center, Worcester, MA, United States*

Learning objectives

After reviewing this chapter, the reader will be able to:

- List the components of a complete urinalysis (physical, chemical, and microscopic examination).
- Discuss methods for evaluating urine appearance and concentration and recognize clinically significant abnormalities.
- Describe methods for identifying the chemical composition of urine.
- List the clinically significant abnormalities identified by microscopic analysis of urine.
- Discuss the limitations of physical, chemical, and microscopic examination techniques in urinalysis and learn the available alternative methods for analysis.

Introduction

Urinalysis is a valuable tool for the clinician. It is most commonly used in the assessment of the genitourinary system; in addition, it can aid in the diagnosis of certain systemic diseases, such as diabetes mellitus or hypertension of pregnancy. Complete urinalysis includes the physical, chemical, and microscopic examination of a urine specimen. Physical examination describes the urine's visual appearance and measured concentration. Chemical analysis is used to reveal and quantify key urine components and is mainly accomplished using dipstick technology. Microscopic evaluation is the third major component of urinalysis; it is the most time-consuming part, yet it remains essential in a number of diagnoses such as urinary tract infections and renal injuries.

Specimen collection and handling

Urine specimens should be collected in a clean, sterile container that has a tightly fitting lid to prevent contamination, evaporation, and spillage. A clean-catch, midstream collection prevents bacteria from the skin of the penis or vagina from contaminating the urine specimen. The sample should be free of vaginal secretions and other extraneous debris. Cleansing of the external genitalia is frequently recommended for women The first morning voided urine is usually the most desirable for urinalysis testing; it is the most concentrated of the day's urine and is also the most acidic (pH between 6.5 and 7.5), so that formed elements such as cells and casts are more stable than in dilute, less acidic urine. Random specimens may be collected at unspecified times and have the advantage of being convenient for patients. They are usually satisfactory for routine screening, although they may not accurately reflect the patient's status, because antecedent factors such as hydration can directly affect urine composition. Timed urine specimens are frequently used to quantitate various aspects of renal function. The urine collected must reflect excretion over a precisely measured duration of time. The specimen must not include urine that is present in the bladder before the timed test begins. The patient should be instructed to obtain a 24-hour urine specimen by discarding the first-voided morning urine on the first day and collecting all subsequent urine *up to and including* the first-voided morning urine on the second day.

Given that the physical, chemical, and microscopic characteristics of a urine specimen begin to change as soon as the urine is voided, urine specimens (except for long-term collections) should ideally be delivered to the laboratory for analysis immediately after collection, which is within 1 or 2 hours of the time of voiding. When this is not possible, steps should be taken to minimize the deterioration of the specimen until it can be transported to the laboratory. Refrigeration is the most commonly used means to preserve urine specimens for delayed analysis; however, it induces precipitation of amorphous or crystalline deposits, which can interfere with the urine microscopic examination. When performing dipstick analysis, the specimen must be allowed to come to room temperature before testing is performed, because low temperatures alter the time required for reactions to come to completion. If urine is allowed to sit unpreserved at room temperature, it will

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00038-7

TABLE 38.1 Changes in delay-analyzed, unpreserved, nonrefrigerated urine specimens.

Physical appearance
• Color may change because of oxidation or reduction of metabolites
• Turbidity may increase because of bacterial growth and possible precipitation of amorphous material
Chemical composition
• pH may increase because of the breakdown of urea to ammonia by urease-producing bacteria
• Glucose may decrease because of metabolic breakdown by patient cells and microorganisms
• Ketones may decrease because of volatilization
• Bilirubin may decrease because of photooxidation
• Nitrite may increase because of bacterial reduction of nitrate
• Bacteria may multiply
• Urobilinogen may decrease because of oxidation to urobilin
Microscopic appearance
• Body cells and casts may disintegrate, especially in dilute, alkaline urine

begin to decompose (Table 38.1). Preservatives work by stabilizing urine so that chemical changes associated with decomposition do not occur and by preventing growth and metabolism of microorganisms. Toluene, phenol, thymol, and acidic or basic preservatives are commonly used for urine chemistry determinations, as is excluding light (bilirubin). Ethanol (95%) or commercially available fixatives may be used to preserve cell structure.

Routine urinalysis

In the kidney, small-molecular-weight substances can freely pass through the glomerulus (see Chapter 35, Laboratory evaluation of kidney function). Most essential components are completely reabsorbed by the tubules; however, increased serum concentration of these substances may exceed the tubules' reabsorptive capacity and lead to their urinary excretion, as with glucose, bilirubin, urobilinogen, and ketones. Many hematological or congenital metabolic disorders can also be detected by increased urinary excretion of particular compounds. Small-molecular-weight or cationic proteins, such as Bence–Jones protein in multiple myeloma, myoglobin in rhabdomyolysis, and hemoglobin in paroxysmal nocturnal hemoglobinuria, may appear in urine through this mechanism. Some substances are normally excreted in urine because of their limited tubular reabsorption; waste products such as urea and creatinine make up most of the normal urinary solutes.

Normally, red blood cells (RBCs) and large-molecular-weight compounds (most serum proteins) are unable to pass through the glomerulus. Therefore the presence of such substances in urine is an indication of an abnormal interface between the bloodstream and the urinary filtrate. While most commonly this indicates the presence of glomerular disease, bleeding also causes blood and protein to be directly released to the urinary tract, which can also indicate nonglomerular diseases such as urinary tract infections, kidney stones, and tumors. As mentioned above, tubular reabsorption prevents the appearance of most small-molecular-weight substances in urine, yet with tubular disease, compounds such as glucose can be found in the urine even with normal blood glucose concentrations, for example, in diabetic nephropathy (see Chapter 35, Laboratory evaluation of kidney function, for more details). Casts and crystals may form directly in the tubular lumen as local concentration of substances is increased by water reabsorption. Such findings can thus point to the presence of intrinsic renal disease. A standard urinalysis historically includes physical examination of the urine, a few simple chemical tests, and microscopic examination of urine sediment.

Physical examination

Appearance

Color

The color of urine is determined to a large degree by its concentration; normal urine varies widely from colorless to deep yellow. Interpretation of color is subjective and varies with each laboratory analyst. It is useful to use

TABLE 38.2 Colors of urine.

	Common causes		
Colors	**Food/drugs**	**Metabolites**	**Clinical conditions**
Yellow to colorless			Polyuria
Yellow	Food color	Urochrome	Healthy patients
Yellow-orange	Warfarin, carotene, and rhubarb	Urobilin and bilirubin	Dehydration
Yellow-green		Bilirubin–biliverdin	Jaundice
Yellow-red-brown	Beets, laxatives (senna), and rhubarb	Hemoglobin, myoglobin, and porphyrin	Hematuria, hemoglobinuria, myoglobinuria, and menstrual contamination
Brown-black	Chloroquine	Porphyrin, melanin, methemoglobin, homogentisic acid, and L-dopa	Alkaptonuria, ochronosis, and rhabdomyolysis
Blue-green	Amitriptyline	Indicant	*Pseudomonas* infection

standardized colored objects in the laboratory as reference points and to use defined colors to describe the urine, avoiding ambiguous terms such as "straw" or "bloody." The clinical significances of common colors of urine are summarized in Table 38.2. Numerous drugs are associated with color changes in urine and should also be considered when interpreting urine color (Table 38.2).

Turbidity

Although cloudy urine is not necessarily pathologic, turbidity in an uncentrifuged urine should be microscopically investigated. Turbidity can be due to the presence of epithelial cells, microorganisms, blood cells, amorphous or crystalline solids, or spermatozoa. Urinary infections and contamination by genital secretions are among the most common causes. Turbidity may also be caused by blood clots, menstrual discharge, pieces of tissue, small calculi, clumps of pus, and fecal material (sometimes from a fistula between the colon or rectum and bladder).

Odor

While the odor of urine is usually given little importance in clinical diagnosis, it can be used to provide clues to certain abnormalities. Normal fresh urine has a faint odor. An offensive, fetid, or ammoniac odor can indicate infection or that a specimen is too old for accurate analysis; a foul odor in a specimen collected more than 2 hours earlier and not preserved or refrigerated indicates an unacceptable specimen. Foul smelling urine in older children and adults is usually due to the presence of infectious agents or accumulation of sulfur-containing compounds in patients with advanced liver disease. In properly collected samples, an ammonia-like odor is suggestive of urea-splitting bacteria, such as *Proteus*, *Pseudomonas*, *Klebsiella*, *Staphylococcus*, and *Mycoplasma*. A few additional characteristic odors can give clues to a medical or congenital disease. Fruity odors indicate the presence of acetone (ketone). A sweet odor is suggestive of the presence of glucose or another sugar, as in diabetes mellitus and maple syrup urine disease. Phenylketonuria is associated with a "mousy" odor and tyrosinuria with a "rotten" odor. Rarely, chemicals derived from food, such as asparagusic acid from asparagus, may be detected.

Concentration

The kidneys control body fluid and electrolyte homeostasis by regulating the volume of excreted water and solutes. Therefore the total urine solute concentration reflects the kidneys' ability to concentrate and dilute urine. This, in turn, is useful in evaluating hydration status. Osmolality and specific gravity both measure the proportion of dissolved solid components in a specimen.

Specific gravity

Specific gravity is a measure of density. More specifically, it is the ratio of the mass of a solution compared with the mass of an equal volume of water. In normal urine, the largest contributors to specific gravity are urea, sodium chloride, sulfate, and phosphate. A number of techniques are available for the measurement of specific gravity. The strip method is an indirect, colorimetric technique using a reagent strip, which contains an electrolyte–dye mixture that elicits a pH change based on the ionic concentration (pKa) of the urine. The reagent-strip

test is fast and simple, and requires no additional equipment. The manufacturer states that there is no interference by glucose, protein, or radiographic dye, all of which interfere with refractometers and urinometers. Disadvantages of reagent strips include their poorer sensitivity than refractometers (units of 0.005), and urine with a pH of 6.5 or greater requiring a correction factor. The refractometer is another indirect method for testing specific gravity. It measures refractive index, which is the ability to bend light, and there is usually a close relationship between the refractive index and the specific gravity. Some large compounds have disparate effects on these two properties and will therefore introduce error. The urinometer is a direct measure of specific gravity and tests buoyancy; however, it is not the method of choice due to the need for larger sample volume and temperature precision. The falling drop method is also a direct measure of specific gravity and is performed by timing a drop of urine and a separate drop of water as they pass through oil. It is more accurate than the refractometer and more precise than the urinometer.

Osmolality

Osmolality measures the total number of dissolved particles per unit of solution (in mmol/L or mOsm/kg water) and reflects the concentrating ability of the kidney better than specific gravity. The normal kidney is capable of producing a urine osmolality ranging from about 40−80 mOsm/kg during diuresis to 800−1400 mOsm/kg during dehydration. In normal urine, over 80% of osmotic activity is due to urea and creatinine, with lesser proportions due to charged solutes such as electrolytes. In normal patients, protein and glucose do not contribute appreciably to urine osmolality. Osmolality is most commonly tested using freezing point depression. For example, 1000 mOsm/kg water depresses the freezing point by 1.86°C.

Chemical examination

A number of chemical constituents are routinely analyzed in urinalysis. The most commonly used techniques include semiquantitative and qualitative reagent strips (dipsticks), tablet tests, and fully quantitative methods for protein, electrolytes, and porphyrins. Some of the factors interfering with urine chemical tests are listed in Table 38.3. The normal ranges of urine components are reviewed in Table 38.4.

Reagent-strip testing

A reagent test strip (or dipstick) is a narrowband of plastic 4−6 mm wide and 11−12 cm long with a linear series of small absorbent pads attached to it. Each pad contains reagents for a separate reaction, so several tests can be carried out simultaneously. Reagent-strip tests have enabled urinalysis laboratories to generate valuable

TABLE 38.3 Interfering factors in urine chemical tests.

Test	Interferences
Specific gravity	Dipstick tests measure ionic strength, which correlates poorly with actual concentration in hospitalized patients; proteinuria and alkaline pH also lead to incorrect results
pH	None
Protein	Alkaline pH may cause false-positives; Bence−Jones protein is not detected
Glucose	Hydrogen peroxide and bleach causes false-positives; ascorbic acid, salicylate, ketones, and L-dopa may cause falsely decreased results
Ketones	Phenylketones, L-dopa, methyldopa, acetylcysteine, captopril, and phthaleins may cause false-positives; ascorbic acid rarely causes false-negatives
Blood	Bleach and microbial peroxides may cause false-positives; ascorbic acid falsely lowers results; high specific gravity or strongly acidic urine can prevent RBC lysis causing a false-negative
Bilirubin	Rifampin and chlorpromazine cause false-positives; exposure to light, ascorbic acid, and nitrite cause false-negatives
Urobilinogen	Many drugs produce false-positives with the Ehrlich method (e.g., sulfonamides)
Nitrite	Contamination may cause false-positives; ascorbic acid may cause false-negatives; Gram-positive bacteria and *Pseudomonas* infections are not detected
Leukocyte esterase	Ascorbic acid, proteinuria, and formalin produce false-negatives

RBCs, Red blood cells.

TABLE 38.4 Composition of urine from healthy subjects.

Constituents	Value (men)	Value (women)
Specific gravity	1.003–1.035 random (1.016–1.022 over 24 h)	
Osmolality	500–850 mOsm/L	
pH	4.6–8.0	
Total protein	< 100 mg/24 h	
Albumin	< 30 mg/24 h or <30 mg/g of creatinine	
Creatinine	20 mg/kg/24 h	15 mg/kg/24 h
Glucose	Usually not detected (<100 mg/dL)	
Ketones	Not detected (<50 mg/dL)	
Total bilirubin	Not detected or ≤ 0.02 mg/dL	
Urobilinogen	< 1 mg	
Urea nitrogen	6–17 g/24 h	
Sodium	Varies with quantity ingested	
Chloride	Similar to sodium excretion	
Potassium	Varies with quantity ingested	
Oxalate	< 45 mg/24 h	
Calcium	<300 mg/24 h	<250 mg/24 h
Uric acid	300–800 mg/24 h	250–750 mg/24 h

semiquantitative chemical results in a rapid, accurate, and efficient manner. In general, properly performed urine test strips are sensitive, specific, and cost-effective. Automated dipstick readers (reflectance photometers) are widely available. To ensure the best results, the urinalysis laboratory should perform the tests according to the following guidelines:

- Test urine promptly at room temperature; use properly timed test readings only.
- Beware of interfering substances.
- Understand the advantages and limitations of the test.
- Employ controls.

Confirmatory testing: tablet and chemical tests

Tablets and chemical tests are sometimes necessary to confirm results obtained by dipstick methods when there are differences in sensitivity (e.g., bilirubin) or specificity (e.g., sulfosalicylic acid for proteins), or to avoid interferences when pigments in the specimens mask the colors obtained on the reagent pads. A confirmatory test should have either the same or better specificity, be based on a different principle, and have a sensitivity equal to or better than that of the original test.

Tests results *often* indicating urinary tract disease

Protein

Healthy individuals will have a daily protein excretion of about 100 mg/day, a very small fraction of the plasma protein content. The normal glomerulus is permeable only to proteins with a molecular weight less than 68,000 Da. Once filtered, proteins are almost completely reabsorbed in the proximal tubule. Proteinuria, therefore, can be the result of either increased or abnormal filtration or decreased reabsorption. Proteinuria, defined as increased protein in the urine, can be classified as transient or persistent. Transient causes include strenuous physical exercise, pregnancy, exposure to extreme cold, relative concentration due to increased water reabsorption, psychological stress, fever, or other acute illnesses. Persistent proteinuria can be subcategorized as glomerular, tubular, or overflow. Glomerular proteinuria is nonselective, and albumin is the predominant urine protein. Tubular proteinuria occurs with tubular injury, resulting in an inability to reabsorb the glomerular filtrate; low-molecular-weight proteins predominate over albumin in this type of proteinuria. Overflow proteinuria occurs when the amount

of low-molecular-weight protein in plasma overwhelms the tubular reabsorptive capacity, as seen with hemoglobinuria, multiple myeloma, and myoglobinuria.

Protein remains the single most useful test on a dipstick for the detection of intrinsic disease of the kidney and urinary tract. Proteinuria is often associated with urine sediment abnormalities, which further aid the diagnosis. Reagent-strip tests are a rapid, inexpensive screening procedure for proteinuria; however, a few important limitations exist. Sensitivity ranges from 6 to 30 mg/dL protein; thus modest increases in urine protein may be missed, including microalbuminuria. In addition, the specificity of these dye-based strip tests is weighted to the detection of albumin, making them unsatisfactory for detection of Bence–Jones proteins, tubular proteinuria, or overload proteinuria. It is recommended that the laboratory communicates these limitations to clinicians and simultaneously performs an alternate method for the detection of all types of protein.

The reagent-strip tests are pH-sensitive. The presence of protein on the strip changes the pH environment of the dye embedded in the pad, resulting in a color change. Highly buffered, alkaline urine can result in a false-positive test.

$$\text{Tetrabromophenol blue} \xrightarrow{\text{pH3}} \text{positive result (green-blue)}$$

$$\text{Tetrabromophenol blue} \xrightarrow{\text{pH3}} \text{negative result (yellow)}$$

A positive or faintly positive result should be confirmed with a more specific test such as the trichloroacetic acid or the sulfosalicylic acid tests. A mildly positive test strip result and a grossly positive turbidity test result may indicate the presence of drugs or Bence–Jones proteins, which can be confirmed on a urine protein electrophoresis.

Blood

The presence of blood in the urine (hematuria) is an important indicator of diseases involving the urinary system. RBCs come from a variety of sources, including the kidneys, ureters, bladder, urethra, prostate gland, uterus, and vagina. Among the more common diseases resulting in hematuria are bladder and kidney infections, tumors, trauma to the kidneys, glomerulonephritis, pyelonephritis, renal calculi, and bleeding disorders related to anticoagulant use. RBCs may be found intact or lysed, which releases hemoglobin and results in hemoglobinuria. Hemoglobinuria, however, is less common than hematuria, and may be associated with trauma, transfusion reactions, severe burns, and poisonings.

From a clinical perspective, hematuria is usually defined by microscopic means, with greater than three RBCs per high-powered field, warranting further workup. The dipstick method provides a low-cost adjunct to microscopy, but its exact role in detecting hematuria is not clearly defined. A positive result deserves further investigation, but a negative result cannot reliably rule out hematuria. The dipstick method detects heme by the peroxidase-like activity of the proteins:

$$H_2O_2 + \text{chromogen} \xrightarrow{\text{Myoglobin or hemoglobin}} \text{oxidized chromogen(blue)} + H_2O$$

Importantly, a positive dipstick test can also indicate hemoglobinuria or myoglobinuria. Myoglobin is a ferrous porphyrin similar to hemoglobin that is commonly seen in urine after crush injuries and muscle trauma and is occasionally seen after intense physical exercise. When myoglobin is released in the circulatory system, it is rapidly excreted by the kidney. The distinction between myoglobinuria and hemoglobinuria must be made using additional tests. Myoglobin may be quantified via immunoassay; however, the presence of markedly elevated serum creatinine kinase is a useful surrogate.

Oxidizing agents such as iodides, bromides, and microbial peroxidases in the urine may cause false-positive results. In menstruating women, a positive result is the rule rather than the exception. False-negative results can occur with high concentrations of ascorbic acid or any compound that can interact with H_2O_2 due to slow down in peroxidase activity. False-negative results may also occur in very concentrated specimens, because the lysing agent may not lower concentration enough to hemolyze the cells and release hemoglobin. False-negative results may also occur in strongly acidic urine specimens, again due to incomplete cell lysis. It may occasionally be important to separate myoglobin from hemoglobin as a cause of a positive result. To confirm hemoglobinuria, the serum from the same patient would be red (hemoglobinemia). In myoglobinuria, the serum would be clear.

Leukocyte esterase

The leukocyte esterase dipstick test is a rapid, inexpensive surrogate for identifying elevated numbers of white blood cells (WBCs) in the urine, that is, pyuria. Common causes of pyuria include pyelonephritis, cystitis, interstitial nephritis, glomerulonephritis, calculi, and corticosteroid and cyclophosphamide (Cytoxan) use. Bacterial infection is the single most common cause of pyuria, and because of this relationship, clinicians may initiate empiric antibiotic treatment in symptomatic pyuric patients. Pyuria is also frequently used to distinguish between true bacteriuria and contamination.

A positive leukocyte esterase test corresponds to approximately 10,000 neutrophils/mL when measured by hemocytometer (see microscopic examination below). An exact definition of pyuria is not universally accepted;

however, a positive leukocyte esterase test is generally considered significant in the proper clinical setting. Leukocyte esterase activity is due to proteins found in primary (azurophilic) granules, which are present in neutrophils, monocytes, eosinophils, and basophils. The reagent strips will detect both lysed and intact leukocytes. The enzymes catalyze the hydrolysis of esters, releasing alcohols and acids that are then used in a color reaction. The intensity of the color is proportional to the number of leukocytes in the specimen.

Most studies have shown good correlation between the presence of neutrophils in the urine sediment examination and a positive leukocyte esterase; however, occasional discrepant results occur, particularly with fewer than 10 WBCs/high-power field (hpf). Leukocyte cell lysis is another cause of discrepant results. Ascorbic acid in very high concentrations may inhibit color development, as will mark proteinuria and aminoglycoside antibiotics. A positive leukocyte esterase (and/or nitrite, see below) occurs in approximately 90%–95% of patients with symptomatic urinary tract infections, and thus warrants a reflex to a urine culture. However, it is much less sensitive in patients with asymptomatic infections, as commonly occurs in the elderly or during pregnancy. Oxidizing agents and formalin preservative may give false-negative results.

Nitrite

This test provides an effective and rapid screening method for detecting the presence of a bacterial infection. The test is based on the principle that most bacteria found in urine have the ability to reduce nitrate, a plentiful constituent of urine, to nitrite. These organisms include *Escherichia coli*, *Klebsiella*, *Proteus*, *Staphylococcus*, and *Pseudomonas*, to mention a few. The determination of clinical bacteriuria is of major significance when culture shows the presence of more than 10^5 colonies/mL. Bacterial infections of the urinary tract are usually accompanied by the presence of large numbers of WBCs in the urine, most often neutrophils. Therefore the nitrite test serves as a valuable screening tool, along with the leukocyte esterase test, in defining whether a urinary tract infection is present. Concomitant negative dipstick results from both tests provide a predictive value of greater than 95% that urine culture will be negative.

The chemical basis of the nitrite reaction is that in an acidic environment, nitrite reacts with an aromatic amine (sulfanilamide or *p*-arsanilic acid) to form a colored diazonium salt that in turn reacts with hydroxybenzoquinoline to provide a pink color (Griess reaction). A positive reaction occurs when the bacterial content is greater than 10^5/mL. It is imperative that fresh urine samples are used when this test is interpreted to eliminate the possibility that the nitrite has been produced by bacterial contamination. First morning voided urine is preferable, because bacteria require up to four hours converting nitrate into nitrite. If bacterial infection is present in the urinary tract, a single dipstick test will be positive in 80%–90% of patients when the first morning specimen is promptly examined. The dipstick nitrite test is qualitative, and any shade of pink produced is considered a positive result, indicative of bacterial infection of the urinary tract.

Negative test results for nitrite are found when the particular species of bacteria does not reduce nitrate. Yeasts and a significant number of nonnitrite-producing, Gram-positive cocci, such as *enterococci*, *streptococci*, or *staphylococci*, may cause infection of the urinary tract and produce a negative nitrite dipstick test. Therefore a negative result should always be verified by urine culture and/or microscopy of the sediment when the patient is suspected of having an infection. False-negative results can also be caused by the presence of high levels of ascorbic acid and urobilinogen, or low pH. Adequacy of diet, especially vegetables, which allow sufficient nitrates to be formed, is essential to the proper interpretation of this test. Antibiotic agents may inhibit the growth of bacteria, even if bacteria are present, and a false-negative result is reported. False-positive results are obtained when various dyes, such as pyridium, discolor the strip test pad pink or red.

Test results *occasionally* indicating urinary tract disease

pH

The kidneys and lungs are the primary regulators of acid–base homeostasis. In healthy individuals, urine pH varies from 4.5 to 8 and most commonly falls between 6 and 6.5. The major pH regulatory mechanisms of the kidney include elimination of hydrogen ions as ammonium, and the reabsorption and production of bicarbonate. Given the kidneys' integral role in acid–base balance, urine pH reflects both renal function and serum pH status. Urine pH is therefore used to recognize certain disease processes, such as renal disease, respiratory disease, and metabolic disorders. It may also be helpful in suggesting the nature of calculi or urine crystals (Table 38.5).

Acidic urine is promoted by dietary consumption of meat protein and some fruits, such as cranberries. Acidic urine is also seen in the cases of metabolic or respiratory acidosis; however, depletion of certain electrolytes, including chloride and potassium, may result in paradoxically acidic urine during systemic alkalosis. Under normal circumstances, the collecting duct reabsorbs chloride in order to eliminate excess bicarbonate, but this function is impaired if the chloride content in the urinary tract is too low, resulting in hypochloremic alkalosis. In hypokalemia, the collecting duct conserves potassium by releasing hydrogen ions in

TABLE 38.5 Crystals in urine.

Objects	Characteristics	pH	Significance
Bilirubin	Brownish red, granular, or needle shape	Acid	Obstructive jaundice
Calcium oxalate	Colorless and octahedral (envelope shape)	Base, neutral	Commonly seen; abnormal in large numbers as with small bowel diseases
Cholesterol	Large, flat, colorless, rectangular, "staircase" pattern, and birefringent	Acid, neutral	Nephrotic syndrome, tumors, and filariasis
Cystine	Flat, hexagonal, and colorless	Acid	Abnormal; seen in inherited reabsorption disorder, renal tubular diseases, and Wilson's disease
Drugs	Dependent on the drug (e.g., sulfadiazine and ampicillin)		Usually high dosage
Leucine	Spherical, brownish yellow, oily-appearing, and refractile	Acid, base	Abnormal; seen in severe liver diseases and maple sugar urine disease
Amorphous phosphate	Fine, granular, birefringent, and white to colorless	Base, neutral	Commonly seen in alkaline urine
Triple phosphate	Birefringent, "coffin-lid" shape, and colorless	Base, neutral	Usually none; often seen in infected urine
Tyrosine	Elongated-needle shaped, colorless to pale yellow, or even black	Acid	Usually abnormal; seen in severe liver disease and inherited metabolic disorders
Uric acid	Various shapes, granular, birefringent, and colorless to pale yellow or reddish brown	Acid	Commonly seen; high numbers seen in tumor lysis and Lesch–Nyhan syndrome, with uric acid calculi

the H^+/K^{+-} ATPase channel, resulting in acidic urine despite systemic alkalosis. Acidic urine may also be induced with medications such as ammonium chloride, methenamine mandelate, and methionine. These are typically used to prevent the formation of certain calculi, namely calcium carbonate, which preferentially forms in alkaline urine.

Alkaline urine may be seen in patients consuming a vegetarian diet or large amounts of citrus fruits. In the cases of metabolic or respiratory alkalosis, alkaline urine is expected. It can also be seen following a large meal, due to gastric acid secretion, a phenomenon commonly known as "postprandial alkaline tide.". Drugs like acetazolamide, potassium citrate, and potassium bicarbonate are used to promote alkaline urine and prevent the formation of certain calculi, namely calcium oxalate, cystine, or uric acid, which preferentially form in acidic urine. Sodium bicarbonate may increase calcium excretion, and therefore is less preferably used in treating calcium stone. Maintaining alkaline urine can also provide a more effective environment for some antibiotic agents.

In patients with renal tubular acidosis (RTA), urine pH may help distinguish the various forms of this condition when combined with other laboratory data. Type 1 RTA is due to impaired hydrogen ion secretion in the collecting tubules, leading to inappropriately elevated urine pH >5.3, despite systemic acidosis. Type 2 RTA is due to impaired bicarbonate reabsorption in the proximal nephron, leading to urinary bicarbonate wasting and elevating the urine pH > 5.3, despite systemic acidosis. (Type 3 is rarely used as a classification, because it is now thought to be a combination of type 1 and type 2.) Type 4 RTA is due to aldosterone deficiency or resistance. This results in a loss of sodium and compensatory reabsorption of potassium. The ensuing hyperkalemia is managed by the H^+/K^{+-} ATPase exchanger, leading to systemic acidosis with appropriately acidic urine, usually pH <5.3.

Measurement of urine pH is made easy by dipstick methodology. The pH pad is impregnated with two indicators, methyl red and bromothymol blue, which change color according to the pH of their environment. The range of blue, green, and orange corresponds to pH between 5 and 9, in half unit increments. To minimize postcollection alteration of pH, the urine should be analyzed promptly and maintained in a tightly sealed container with minimal headspace. Open containers allow gas exchange and the release of carbon dioxide, which increases pH. Urine pH may also increase due to the growth of bacteria and breakdown of urea.

Sugars

Glucose is one of a variety of reducing substances that may be present in the urine, and from a clinical

standpoint, it is of particular interest due to its various disease associations. Small amounts of glucose may be present transiently in the urine of healthy patients, especially after ingesting foods or beverages that have a high concentration. Glucose is freely filtered by the glomerulus and almost entirely reabsorbed by the renal tubules, so that under normal circumstances, little is found in the urine. Glycosuria is defined as the presence of detectable glucose in the urine. This generally occurs when blood levels of glucose exceed 180 mg/dL. At these levels, with similar concentrations occurring in the glomerular filtrate, the renal threshold for reabsorption is exceeded, and any excess glucose is excreted in the urine.

In diabetes mellitus, serum concentrations of glucose may reach very high levels, and glucose may be seen in the urine. Glucose may also be found in the urine in other pathological conditions such as those affecting the central nervous system (stroke and neoplasms), kidneys (uremia), endocrine system (overproduction of adrenocorticotropic hormone), and liver (glycogen storage disease) or may be related to general metabolic problems, such as feeding after starvation. Some pharmaceutical agents, such as diuretics and birth control pills, may also cause glycosuria.

Reducing sugars other than glucose are occasionally seen in urine in patients with inherited metabolic disorders. Galactosemia is one such disorder. Deficiency of galactose-1-phosphate uridyltransferase (GALT) or galactokinase is associated with developmental delay and other sequelae that may be limited with early dietary restriction of galactose. Congenital lactase deficiency results in lactosuria, malabsorption, and diarrhea starting at birth; management with lactose-free formula is life-saving. Sucrase deficiency is less common than lactase deficiency but has similar symptoms; it results in sucrosuria. Other sugars that may be present in the urine are fructose and pentose, both infrequent causes of nonglucose mellituria.

Glucose can be measured in various ways. Reagent strips employ the specific enzyme activity of glucose oxidase. The enzyme catalyzes the conversion of glucose into gluconic acid and hydrogen peroxide. A second enzyme, peroxidase, promotes the reaction of hydrogen peroxide with a chromogen (e.g., iodine complex) to form a brown oxidized compound. The chromogen varies with the manufacturer, as does the sensitivity.

$$\text{Glucose} + O_2 \xrightarrow{\text{Glucose oxidase}} \text{gluconic acid} + H_2O_2$$

$$H_2O_2 + \text{chromogen} \xrightarrow{\text{Peroxidase}} \text{oxidized colored chromogen} + H_2O$$

Results of the glucose dipstick are semiquantitative and reported in a 1 + to 4 + format, which are equated to concentrations ranging from approximately 100−2000 mg/dL. Other sugars do not give a positive dipstick test. False-positive results occur when oxidizing agents act directly on the chromogen; these include bleach and microbial peroxides. False-negative results may occur when reducing agents like ascorbic acid prevent the oxidation of the chromogen. False-negative results may also occur when glucose has been metabolized by bacteria in samples that have been allowed to stand after voiding.

The copper reduction tablet or Benedict's test (Clinitest, Bayer Diagnostics) provides an additional method for testing urine sugar. In addition to glucose, Benedict's test will detect the presence of all reducing sugars, including galactose, lactose, fructose, maltose, and pentoses. It is, therefore, an important method for use in pediatric patients to screen for disorders of carbohydrate metabolism. The goal of this screen is primarily to detect galactosemia, as other disorders of carbohydrate metabolism are rare and generally more clinically benign. Sucrose is not a reducing sugar and will not be detected. To verify the presence of a nonglucose-reducing sugar, comparisons should be made with a dipstick result from the same patient.

The copper reduction method is performed in a test tube by adding urine and water to a tablet composed of copper sulfate, sodium citrate, sodium carbonate, and sodium hydroxide. The sodium carbonate and citric acid act as effervescents, rapidly dissolving the tablet. Sodium hydroxide creates an alkaline environment and reacts with citric acid and water to produce heat. In this environment, the cupric ion reacts with reducing substances in the urine to produce cuprous ions. Shortly after the components are combined, if reducing sugars are present, red color will develop in the solution. The bottom of the test tube will be hot and should not be touched. The amount of sugar present is then quantitated by comparing the color of the products with a color chart.

$$\text{Cupric ions} + \text{glucose (or reducing substances)} \xrightarrow{\text{Heat and alkali}} \text{cuprous oxide (red)}$$

The Clinitest tablet will detect reducing substances at a concentration of 250 mg/dL or greater. Clinitest is both less specific and less sensitive than the reagent strip; it cannot be used as a confirmatory test for a positive reagent-strip glucose test. False-positive results occur in the presence of ascorbic acid and certain drugs, such as nalidixic acid, probenecid, and cephalosporins and in patients with kidney injury and/or liver disease. Normal newborns may excrete sugars such as glucose, galactose, fructose, and lactose during the first 10−14 days of life. Normal pregnant and postpartum women may excrete lactose. For these reasons, screening for carbohydrate metabolism disorders by the Bayer Clinitest reagent tablets, which have been discontinued by the manufacturer, or any other similar method, is no longer recommended. Urine tests for carbohydrates can be used in lieu of this test as well.

Rather than detecting reducing substances in urine, the current recommended method uses a blood specimen and

detects the activity of the deficient enzyme in galactosemia (GALT). This method has improved sensitivity and specificity compared with the previously used urine test. Thin-layer chromatography can also be utilized in the specific identification of nonglucose-reducing sugars when inherited inborn metabolic diseases of sugar metabolism are considered.

Test results *rarely* indicating urinary tract disease

Ketones

Ketone production is a result of metabolic alterations that occur in the setting of inadequate carbohydrate supply. Normally, the body completely breaks down fatty acids into carbon dioxide and water. However, when the supply of carbohydrate is poor, there is an increase in lipolysis. Some of this fat is metabolized into ketones using the Krebs cycle intermediate, acetyl-coenzyme A. Through sequential reactions, acetyl-coenzyme A is converted into acetoacetic acid, acetone, and beta-hydroxybutyric acid, which are collectively referred to as "ketone bodies." These three products appear in the blood (ketosis) and urine (ketonuria), with acetone accounting for only 2%, acetoacetic acid for 20%, and beta-hydroxybutyrate making up the remainder. Acetoacetic acid and beta-hydroxybutyrate are acids and can contribute to the development of ketoacidosis.

Diabetes mellitus is the most common cause of ketoacidosis, and screening for urinary ketones is an important part of disease monitoring. Patents with type 1 diabetes are at higher risk of developing ketosis and ketoacidosis than are type 2 diabetics. Nondiabetic causes of ketonuria include low carbohydrate dieting, cachexia, anorexia nervosa, extreme calorie deprivation, and acute illnesses accompanied by vomiting and diarrhea. An inherited metabolic disorder should be considered in infants with severe, prolonged ketonuria. In pediatric patients, ketonuria can be caused by an acute febrile illness.

Dipstick testing for ketone bodies is based on the principle that both acetone and acetoacetic acid react with sodium nitroprusside in an alkaline buffer solution to produce a purple complex (Legal's test, Chemstrip). Some reagent strips (Multistix) do not contain alkali, and thus do not react with acetone. Beta-hydroxybutyric acid does not react with either strip formulation. The lower limit of detection for reagent-strip testing is approximately 5 mg/dL acetoacetic acid and 70 mg/dL of acetone.

$$\text{Acetoacetic acid} + \text{sodium nitroprusside} + \text{glycine} \xrightarrow{\text{AlkalinepH}} \text{purple color}$$

Reagent strips correlate only moderately well with serum ketones. False-positive results can be due to dyes, large amounts of phenylketones, acetylcysteine, methyldopa, captopril, L-dopa metabolites, or the preservative 8-hydroxyquinolone. Reagents are labile, and false-negative results can occur due to the deterioration of reagents.

Nitroprusside tablet tests (Acetest, Bayer Diagnostics, discontinued) also exist for the detection of urinary ketones and can be used on serum samples as well. The color reaction is slightly different and may be helpful when the urine has an interfering color. The lower limit of detection is 5 mg/dL acetoacetic acid and 20 mg/dL acetone. Beta-hydroxybutyrate is not detected. Sources for false-positives and negatives are otherwise identical to those seen in the dipstick method. Since neither the dipstick nor the tablet method detects beta-hydroxybutyrate, a serum test was developed for a quantitative result.

Bilirubin

Bilirubin is a breakdown product of hemoglobin and is formed in the reticuloendothelial cells of the liver, spleen, and bone marrow. It is water-insoluble and travels through the bloodstream bound to albumin (termed indirect or unconjugated bilirubin). Unconjugated bilirubin cannot be excreted by the kidneys, because albumin is not normally filtered through the glomerulus. Even in the cases of renal failure, the quantity of albuminuria and bilirubinemia is typically not sufficient to reach a detectable range of bilirubin in the urine.

Within the liver, bilirubin is chemically conjugated to glucuronic acid, which creates a water-soluble form (termed direct or conjugated bilirubin). Conjugated bilirubin is freely filtered across the glomerulus. In normal circumstances, however, most conjugated bilirubin never reaches the kidney, because it is secreted into bile. Normal urine contains small quantities of conjugated bilirubin ($\sim$0.02 mg/dL), but concentrations this low are not detectable using ordinary dipstick methodology. Therefore the presence of conjugated bilirubin in the urine usually reflects a pathologic increase in serum bilirubin. This can occur in two general circumstances, either biliary obstruction or hepatocyte dysfunction resulting in an inability to excrete conjugated bilirubin into bile. Causes of biliary obstruction include gallstones in the common bile duct, pancreatic carcinoma, drug-induced cholestasis, and periportal inflammation and fibrosis. Causes of hepatocyte dysfunction include acute hepatitis due to viruses, drugs, alcohol, and other toxins, as well as congenital Dubin–Johnson and Rotor syndromes. Bilirubinuria may occasionally be diagnosed before clinical jaundice appears. One crude but rapid method for the determination of bilirubinuria is vigorous shaking of a urine-filled test tube. Shaking produces yellow foam in the presence of increased quantities of bilirubin, whereas urine with a normal amount of bilirubin will produce white foam.

Dipstick methodology for the semiquantitative determination of bilirubinuria utilizes a reaction between the bilirubin and a diazotized aniline dye, resulting in a colored azobilirubin compound. The exact aniline dye used in the dipstick and the resulting color varies with the manufacturer. Results are interpreted as negative (<0.5 mg/dL) to 1+ (small amounts of bilirubin) to 3+ (large amounts of bilirubin). Ictotest reagent tablets (Bayer Diagnostics, discontinued) can serve as a confirmatory test. The tablets utilize similar diazotized dyes but are more sensitive (as low as 0.05 mg/dL) than the dipstick pads and less affected by interfering substances.

$$\text{Bilirubin glucuronide} + \text{diazonium salt} \xrightarrow{\text{acid}} \text{azobilirubin (colored)}$$

It is important that the urine be tested for bilirubin immediately after voiding, because bilirubin is unstable when exposed to light and breaks down rapidly. The conjugated form is quickly hydrolyzed into free bilirubin, which is less reactive. Prolonged exposure to light will cause the dipstick pad to become negative, owing to the oxidation of bilirubin to biliverdin, a substance that is nonreactive to diazo dyes. False-negative results are also obtained with the dipstick method when the tested urine contains large quantities of ascorbic acid or nitrites (released because of bacterial growth). False-positives may be due to colored drugs and their metabolites, including rifampin, chlorpromazine, and phenazopyridine. The tablet method is less affected by the hydrolysis of the glucuronide form and is not affected by ascorbic acid or nitrites.

Urobilinogen

Urobilinogen is also a product of hemoglobin metabolism and is formed by bacterial degradation of bilirubin in the colon. Approximately one half of urobilinogen is reabsorbed into circulation from the intestine. This reabsorbed urobilinogen is mostly removed from portal venous circulation and reexcreted into the bile. Only a small amount of urobilinogen typically reaches systemic circulation, amounting to a urinary excretion of no more than 4 mg per day. A random urine sample may be expected to contain 0.1−1 mg/dL urobilinogen.

The level of urobilinogen in the serum and urine is useful in narrowing the clinical differentials for jaundice. In the absence of obstruction, urobilinogen levels generally rise with increased bilirubin production. If bilirubin does not reach the lower gastrointestinal (GI) tract, however, as occurs with obstruction of the upper GI tract or common bile duct, urobilinogen is not produced and will not be detected in the urine. Elevated urobilinogen levels may also be seen when the liver simply cannot remove urobilinogen from the portal venous blood, as occurs in severe liver disease, such as cirrhosis.

Two dipstick methods for detecting urobilinogen are available and are sensitive to concentrations greater than 0.2−0.4 mg/dL. In the Ehrlich aldehyde reaction (Multistix), urobilinogen is mixed with dimethylaminobenzaldehyde in an acid buffer, resulting in a tan to orange-colored compound. The Ehrlich method is not specific for urobilinogen and will react with porphobilinogen, sulfonamides, procaine, indole, methyldopa, 5-hydroxyindoleacetic acid, and aminosalycylic acid metabolites. The Ehrlich reaction may be inhibited by high concentrations of nitrites, produced in bacterial metabolism, resulting in a false-negative result. The newer Chemstrip method is specific for urobilinogen. In this test, urobilinogen reacts with a diazonium compound in an acid medium, producing a red azo dye. Both methods are affected by drugs and metabolites that mask the color of the test strip and lead to erroneous interpretations; these include phenazopyridine and drugs containing azo dyes. False-negative results using either methodology can be obtained when testing is delayed, since urobilinogen can be converted into urobilin, which is nonreactive. False-negative results can also occur for both methods if formaldehyde is used as a urine preservative, since it will inhibit the reaction.

Microscopic examination

The microscopic evaluation of urine sediment is utilized for the early recognition of infectious, inflammatory, and neoplastic conditions of the urinary system. The ideal specimen for microscopic examination is a first morning voided urine; however, a randomly collected sample is likely satisfactory for most screening purposes. If a sample cannot be examined within 2 hours, it should be preserved, but one must be familiar with the limitations of the preservation technique. Refrigeration minimizes bacterial growth and associated metabolic changes to the sample, but it may result in the formation of crystals or other amorphous elements. Formalin preserves cellular details but can interfere with dipstick and other chemical tests.

Following proper sample collection, the urine may be concentrated using centrifugation, followed by slide preparation. Each laboratory uses different methods, which means that each must establish its own reference values for microscopic examination. Standardization of microscopic urinalysis is essential to reduce ambiguity and minimize subjectivity. Aspects of the microscopic examination that should be standardized are:

1. volume of urine analyzed,
2. duration and force of centrifugation,
3. resuspension volume and concentration of sediment,
4. volume and amount of sediment examined, and

5. terminology and reporting format.

A number of techniques are available for sediment examination. Standardized bright-field microscopy is still the most common technique employed. Supravital staining can be combined with bright-field microscopy to enhance cellular details. Typically, a small quantity of the supravital stain (e.g., crystal violet and safranin) is added to the concentrated sediment. Polarizing filters are also used to supplement bright-field microscopy and can distinguish certain morphologically similar elements. Phase-contrast microscopy is probably the best method for rapid urine sediment evaluations without the use of stains and is particularly useful in visualizing translucent objects.

Quantification of formed elements is commonly accomplished using a hemocytometer chamber. The chamber contains a known volume of sample and a grid. The formed elements within a portion of the grid are counted and converted into an overall sample concentration. Commercially available standardized slide methods are a practical alternative to the hemocytometer and are far superior to the conventional glass slide and coverslip method. Systems to automate partially the microscopic urinalysis are utilized by some laboratories and are becoming increasingly sophisticated. One method uses a laminar flow chamber to detect various elements within a given volume and present them as images that can be sorted and identified by the user. Flow cytometry is mainly utilized for counting cells and casts.

Crystals

Crystalline forms are not normally seen in the freshly voided urine sediment; however, precipitation of crystals will occur during refrigeration or when the urine has been allowed to stand. There are only a few types of clinically significant crystals, and their proper identification is often reliant upon both microscopic examination and pH at the time of crystal formation. The pH of normal urine varies from acid to alkaline, and each crystal type is associated with a specific pH range. The formation of crystals is also related to the concentration of various salts, which, in turn, depends on the patient's metabolic state, diet, and hydration status. Crystals seen in normal alkaline urine include amorphous phosphates, triple phosphates, and calcium carbonate. Normal acidic urine may also include small quantities of crystals, including uric acid and calcium oxalate. Most such normally occurring crystals have little if any pathologic significance. Crystals and associated disease states are summarized in Table 38.5.

Blood cells

Red blood cells

Hematuria accompanies a number of clinical conditions (see chemical analysis of hematuria above) and is typically defined as three or more RBCs per hpf. Microscopic analysis of the sediment is essential to the clinical categorization of hematuria. Three major categories exist and serve to locate the origin of the RBCs, allowing the clinician to narrow the differential diagnosis. Glomerular hematuria indicates a glomerular source for the RBCs and is distinguished by the presence of significant proteinuria, dysmorphic RBCs, and erythrocyte casts. The most common cause of glomerular hematuria is IgA nephropathy. Nonglomerular hematuria presents with significant proteinuria but no dysmorphic RBCs or erythrocyte casts. This pattern is due to tubulointerstitial, renovascular, or metabolic diseases. Urologic hematuria is hematuria without proteinuria, dysmorphic RBCs, or erythrocyte casts. Associated diseases include tumors, infection, and calculi.

RBCs that enter the urine proximal to the tubules are more likely to become deformed by passage through the highly concentrated loop of Henle, leading to dysmorphic RBCs and enabling the observer of such cells to deduce a glomerular origin. These dysmorphic RBCs include any deviation from the normal RBC morphology but most commonly include acanthocytes. Erythrocyte casts are indicators of glomerular bleeding. They can rarely be seen in tubulointerstitial disease and have also been identified in healthy individuals participating in contact sports.

White blood cells

In the centrifuged urinary sediment, WBCs are normally present in relatively small numbers, usually fewer than five per hpf. Increased numbers of leukocytes in the urine represent inflammation of the urinary tract and can be due to infectious or noninfectious causes. Localized infections are termed cystitis, pyelonephritis, urethritis, prostatitis, or balanitis. Staphylococcal and enteric bacteria are the most common infectious agents. Noninfectious causes of pyuria include interstitial nephritis, glomerulonephritis, systemic lupus erythematosus, tumors, calculi, fever, certain drugs, and even strenuous exercise.

While neutrophils are by far the predominant leukocyte present in urine, other WBC types may also be seen. An increased proportion of lymphocytes and mononuclear cells is suggestive of a chronic infection, while eosinophils are an important indicator of acute interstitial nephritis. WBC casts are another important diagnostic element. Their presence indicates renal involvement, most commonly due to bacterial pyelonephritis.

Epithelial cells

Renal tubular epithelial cells

RTE cells originate in the proximal and distal convoluted tubules and are best identified using the Papanicolaou stain. They are large, oblong cells with granular cytoplasm,

which can make it difficult to distinguish them from granular casts. The nucleus is small, dense, and often eccentric. Multinucleation can be seen. RTE cells are normally found in the urine in small numbers because of physiologic sloughing and regeneration. Slightly larger numbers of these cells can be seen in normal newborns but are otherwise considered pathologic. The specific etiologies for increased numbers of RTE cells in the urine are numerous and include acute tubular necrosis, acute ischemic injury, heavy metal poisoning, mushroom poisoning, and drug toxicity, particularly aminoglycosides, nonsteroidal antiinflammatory drugs, and immunosuppressants. Cytoplasmic inclusions may also be present and include lipid, hemosiderin, bilirubin, and melanin. Lipid inclusions are called oval fat bodies and support a diagnosis of nephrotic syndrome when accompanied by proteinuria.

Collecting duct epithelial cells

Collecting duct cells are cuboidal, columnar, or polygonal but never round. They have a large, slightly eccentric nucleus. Increased numbers are seen in renal transplant rejection, acute tubular necrosis, ischemic insults, malignant nephrosclerosis or acute glomerulonephritis with tubular damage, as well as drug or chemical injuries. Salicylate intoxication is included among these. A cluster of three or more collecting duct cells is called a renal epithelial fragment and is indicative of ischemic necrosis. The presence of renal epithelial fragments is frequently accompanied by pathologic casts and reflects a more severe injury with basement membrane disruption.

Urothelial cells

These are frequently present in the urine sediment (0−1 per hpf) owing to their high turnover rate. The cells are smaller than squamous cells and have a round, centrally located nucleus, small nucleoli with abundant homogeneously granular, or finely vacuolated cytoplasm. The cytoplasm is usually round or pear-shaped, and occasionally has a tail-like projection. Increased numbers of urothelial cells can be seen in certain inflammatory conditions affecting the urinary tract. Large numbers of polymorphonuclear neutrophils, plasma cells, and lymphocytes are also indicative of inflammation. Urothelial cells are rarely of pathological significance except when they are seen in large groups or sheets in the absence of instrumentation. Clusters of atypical urothelial cells can be seen in urothelial carcinoma. In such cases, an additional sediment sample should be obtained for cytologic examination.

Squamous cells

Squamous cells are easy to recognize, because they are large, flat, contain a small central nucleus, and frequently encountered. In a well-collected, midstream, "clean-catch" specimen, there should be fewer than 1−2 per hpf. While some squamous cells can be exfoliated and picked up from foci of squamous metaplasia in the trigone of the bladder (especially in women), larger numbers of squamous cells in the urine could also represent cervicovaginal/genitourinary contamination. One special type of squamous cell is the "clue" cell, an epithelial cell with a distinctive stippled or cloud-like appearance owing to the numerous bacteria coating the cell surface. Clue cells are a sign of bacterial vaginosis, which is caused by *Gardnerella vaginalis*, a Gram-negative coccobacillus. These bacteria displace the normally seen lactobacilli, which are Gram-positive rods, and cause inflammation, increased pH, and foul smelling vaginal discharge.

Casts

Casts are formed when a particular component in urine solidifies during passage through the tubule. For example, Tamm−Horsfall protein is a high-molecular-weight protein that is secreted by tubular epithelial cells. As water is reabsorbed, the concentration of this protein increases; under extreme circumstances, this leads to the formation of a gel that partially solidifies and takes the shape of the tubule in which it formed, termed a cast. The Tamm−Horsfall-derived cast matrix is either hyaline or waxy; waxy casts are a denser, highly refractile matrix. Factors that enhance cast production include increased protein concentration, increased ionic concentration, lower pH, and stasis or obstruction of the nephron by cells or cellular debris. Even a normal amount of urine protein excretion may precipitate when urine becomes highly concentrated; hence, casts are a common feature in extreme dehydration, after exercise, and during fever. Additional proteins contributing to cast formation include Bence−Jones as well as albumin; the latter may be seen in glomerular injury or extreme exercise.

The size and shape of casts indicate the condition of the tubules where the precipitation occurred; thus wide casts occur in dilated tubules (as seen in patients with renal failure), convoluted casts develop in the convoluted tubules, etc. When a cast forms, it traps any cells that are also present in the urine; thus casts serve as a "biopsy" of the environment within the kidney. The classification of renal casts is simple and based on morphologic appearance. In the normal state, only two varieties of casts appear in the urinary sediment: hyaline casts and finely granular casts. Any additional cast forms must be considered abnormal and are most often associated with generalized metabolic or intrinsic renal disease. Casts, in contrast to cells, are quantitated in the sediment as the number per low-power microscopic field. Common urinary casts are summarized in Table 38.6.

TABLE 38.6 Common renal (urinary) casts.

Types	Characteristics	Significance	Associated disease(s)
Physiologic or pathologic			
Hyaline	Transparent	Exercise, dehydration, and fever	Nonspecific
Granular	Semitransparent containing refractile granules representing protein, salts, or other cellular material	Exercise, dehydration, fever, and glomerular and tubular disease	Nonspecific
Pathologic			
Cellular			
Erythrocytic	Semitransparent or granular cylinder containing distinct erythrocytes	Renal parachymal bleeding, glomerular leakage	Acute glomerulonephritis, IgA nephropathy, lupus nephritis, and renal infarction
Blood	Yellow to red granular cylinder but intact erythrocyte stroma not seen	The same as above	The same as above
Leukocytic	Transparent granular or waxy cylinder containing segmented neutrophils	Interstitial inflammation	Tubulointerstitial disease (pyelonephritis), glomerular disease, interstitial nephritis, and lupus nephritis
Renal tubular epithelial	Semitransparent granular or waxy cylinder containing intact or necrotic renal tubular epithelial cells	Tubular damage	Acute tubular necrosis, acute allograft rejection, tubulointerstitial disease, viral disease, and drugs
Noncellular			
Waxy	Sharply defined, broad, with broken off borders and indentations, homogeneous, and highly refractile	Cellular degeneration, tubular inflammation, and extreme urinary stasis	Nonspecific; seen in many disease processes
Fatty	Semitransparent or granular cylinder containing large, highly refractile vacuoles or droplets	Lipiduria	Nephrotic syndrome
Bilirubin	Deep yellow to brown, transparent, granular waxy cylinder	Elevated conjugated bilirubin	Obstructive jaundice and liver dysfunction
Crystal	Crystalline inclusion in a semitransparent or granular cylinder	Deposition of crystals in the tubule or collecting duct	Nonspecific
Other			
Broad	Width of cylinder two to six times that of other casts; most commonly waxy or granular types	Tubular dilation and stasis	Advanced renal disease

Microorganisms

In a properly collected and processed urine specimen, the presence of microorganisms is clinically significant. The identification of bacteria on an uncentrifuged specimen using Gram staining is suggestive of greater than 10^6 organisms/mL, which is considered significant. Bacteria is best quantified and identified through the microbiology laboratory, where the presence of 10^5 colony forming units per mL or greater is usually associated with a urinary tract infection.

Fungi may also be identified in the urine sediment. *Candida* species are the most common among these and are typically observed in the yeast form. As with bacteria, the presence of fungi may represent contamination from the female genital tract or the environment. Parasites such as *Trichomonas vaginalis* or parasitic ova may also be seen. Since these organisms are more commonly observed due to genital or fecal contamination, a repeat clean-catch urine specimen is recommended. Finally, viral cytopathic effects may be observed in the epithelial cells of the urine sediment and are more commonly seen in immunosuppressed patients.

Some of the more common causative viral agents include herpesviruses, cytomegalovirus, and polyomavirus. Cytologic techniques should be used for accurate identification.

Automated urinalysis platforms

Automated urinalysis platforms have been introduced in the last decade. Some have a urine chemistry analyzer integrated with an automated microscopy analyzer on a common sample track (i.e., Beckman Iris system), and some are desktop point-of-care urinalysis strip readers (i.e., Siemens CLINITEK Status system). The point-of-care urinalysis strip readers can now perform protein-to-creatinine ratio or albumin-to-creatinine ratio for diabetes or chronic kidney disease and human chorionic gonadotropin pregnancy test, which would be extremely valuable in the Emergency Department. With the expansion of test menu and automated technology, urinalysis can be more accurate, reliable, and efficient.

Suggested reading

[1] J.A. Simmerville, W.C. Maxted, J.J. Pahira, Urinalysis: a comprehensive review, Am. Fam. Physician 71 (2005) 1153–1162.

[2] L.J. McBride, Textbook of Urinalysis and Body Fluids, Lippincott-Raven, Philadelphia, PA, 1998.

[3] L.A. Kaplan, A.J. Pesce, Examination of urine, in: L.A. Kaplan, A. J. Pesce, S. Kazmierczak (Eds.), Clinical Chemistry: Theory, Analysis, Correlation, fourth ed., Mosby, St. Louis, MO, 2003, pp. 1092–1109.

[4] M.H. Haber, P.C. Ward, Urine, in: K.D. McClatchey (Ed.), Clinical Laboratory Medicine, second ed., Lippincott Williams & Wilkins, Philadelphia, PA, 2002, pp. 519–551.

[5] D.R. Dufour, Urinalysis, in: D.R. Dufour (Ed.), Professional Practice in Clinical Chemistry: A Companion Text, AACC, Washington, DC, 1999. 26-1-15.

[6] R.A. McPherson, J. Ben-Ezra, S. Zhao, Basic examination of urine, in: R.A. McPherson, M.R. Pincus (Eds.), Henry's Clinical Diagnosis and Management by Laboratory Methods, twenty-first ed., WB Saunders, Philadelphia, PA, 2007, pp. 390–454.

[7] W. Clarke, D.E. Palmer-Toy, Clinical chemistry of urine, pleural effusions, and ascites, in: K.D. Lewandrowski (Ed.), Clinical Chemistry: Laboratory Management and Clinical Correlations, Lippincott Williams & Wilkins, Philadelphia, PA, 2002, pp. 849–864.

[8] C.Y. Pak, K. Sakhaee, C. Fuller, Successful management of uric acid nephrolithiasis with potassium citrate, Kidney Int. 30 (1986) 422–428.

Self-assessment questions

1. What are the components of a complete urinalysis?
   - **a.** Concentration, chemical examination, and microscopic examination
   - **b.** Physical examination, dipstick analysis, and microscopic examination
   - **c.** Concentration, dipstick analysis, cell, cast, and crystal analysis
   - **d.** Physical examination, chemical examination, and microscopic examination
2. Which of the following can occur due to delayed analysis of an unpreserved, unrefrigerated urine specimen?
   - **a.** Decreased bilirubin, urobilinogen, ketones, glucose, and nitrites
   - **b.** Increased bacterial count, turbidity, pH, ketones, and nitrites
   - **c.** Decreased bilirubin, urobilinogen, ketones, glucose, pH, and casts
   - **d.** Increased bacterial count, turbidity, pH, ketones, and darkened color
3. Which of the following associations is incorrect?
   - **a.** Red urine due to beets and rhubarb ingestion
   - **b.** Brown-black urine due to methemoglobin and porphyrin
   - **c.** Orange urine due to bilirubin and urobilinogen
   - **d.** Green urine due to pseudomonas infection
4. Which of the following associations is incorrect?
   - **a.** Mousy odor due to phenylketonuria
   - **b.** Sweet smell due to maple syrup urine disease
   - **c.** Fruity odor due to tyrosinuria
   - **d.** Fetid odor due to urinary tract infection
5. A clinically hydrated patient is found to have a urine specific gravity of 1.050. Which of the following potentially explains this scenario?
   - **a.** The patient recently received radiocontrast media.
   - **b.** This is a normal result for a hydrated patient.
   - **c.** The patient has diabetes insipidus.
   - **d.** The patient has acidic urine.
6. Which method of protein detection is not able to detect immunoglobulin light chains (i.e., Bence–Jones proteins) in urine at a level >10 mg/dL?
   - **a.** Immunoelectrophoresis
   - **b.** Reagent-strip protein test
   - **c.** Sulfosalicylic acid precipitation test
   - **d.** Coomassie blue dye-binding colorimetric method
7. Reagent-strip testing for hematuria is expected to be positive in which of the following circumstances?
   - **a.** Strongly acidic urine
   - **b.** Crush injury with myoglobinuria
   - **c.** High concentrations of ascorbic acid
   - **d.** High urine specific gravity
8. Leukocyte esterase is _____.
   - **a.** a protein found in primary granules of many leukocytes
   - **b.** a dipstick test that delivers false-positive results with the use of formalin-preserved specimens
   - **c.** a bacterial enzyme that catalyzes the hydrolysis of esters, releasing alcohols and acids

d. a dipstick test that is only activated by lysed leukocytes

9. Which of the following dipstick tests rely upon formation of a colored diazonium salt and, therefore, react with significant amounts of nitrite?
a. Glucose
b. Ketones
c. Blood
d. Bilirubin

10. Two methods are used to assess the presence of glucose in a urine specimen. The reagent test strip impregnated with glucose oxidase–peroxidase reagent yields a negative result, but Benedict's test yields a positive result. What is the best explanation?
a. Benedict's test is more sensitive than the strip method.
b. A nonglucose, reducing substance is present in the urine sample.
c. The urine sample has been contaminated by hypochlorite.
d. The urine sample has been contaminated by peroxide.

11. A routine urinalysis is performed on a young child suffering from diarrhea. The reagent test strip is negative for glucose but positive for ketones. These results may be explained by which of the following statements?
a. The child has Type I diabetes mellitus.
b. The child is suffering from lactic acidosis, and the lactic acid has falsely reacted with the impregnated reagent area for ketones.
c. The child is suffering from increased catabolism of fat because of decreased intestinal absorption.
d. The reagent area for ketones was read after the maximum reading time allowed.

12. Ascorbic acid can cause false-negative results for which reagent-strip tests?
a. Blood
b. Ketones and glucose
c. Leukocyte esterase
d. Nitrite
e. All of the above

13. A urine specimen tests positive for bilirubin but negative for urobilinogen. Both tests use a diazonium-based method. What is the most likely clinical scenario?
a. Obstruction of the common bile duct
b. Cirrhosis
c. Acute viral hepatitis
d. Administration of rifampin

14. The identification of unstained cellular components and casts in urine sediments provides a practical method for the detection and differentiation of formed elements in urinary sediment. Which microscopic technique is used for this procedure?
a. Fluorescent microscopy
b. Phase-contrast microscopy
c. Polarized microscopy Bright-field microscopy

15. Oval fat bodies, fatty, waxy renal tubular casts, and cholesterol crystals are seen in which of the following diseases?
a. Nephrotic syndrome
b. Chronic glomerulonephritis
c. Acute renal allograft rejection
d. Chronic pyelonephritis

Answers

1. d
2. d
3. b
4. c
5. a
6. b
7. b
8. a
9. d
10. b
11. c
12. d
13. a
14. b
15. a

Chapter 39

Disorders of the anterior and posterior pituitary

Nichole Korpi-Steiner[1] and Robert D. Nerenz[2]

[1]*Department of Pathology and Laboratory Medicine, University of North Carolina, Chapel Hill, NC, United States,* [2]*Department of Pathology and Laboratory Medicine, Dartmouth-Hitchcock Health System, Lebanon, NH, United States*

Learning objectives

After reviewing this chapter, the reader will be able to:

- List the anterior and posterior pituitary hormones, their sites of action and expected physiological effects.
- Define the regulatory mechanisms governing the synthesis and release of the anterior and posterior pituitary hormones.
- Describe clinical signs and symptoms consistent with dysregulated pituitary hormone function.
- Recommend appropriate laboratory tests for the initial evaluation of suspected pituitary dysfunction and recognize patterns of test results associated with each condition.
- Suggest provocative testing to identify the mechanism responsible for apparent pituitary dysfunction.

Introduction

The pituitary gland (hypophysis) is composed of two anatomically and physiologically distinct portions: the anterior pituitary (adenohypophysis) and the posterior pituitary (neurohypophysis). The anterior pituitary is frequently referred to as the endocrine "master gland" due to its central role in coordinating many essential endocrine systems. Anterior pituitary hormones are responsible for maintaining reproductive health [luteinizing hormone (LH) and follicle-stimulating hormone (FSH)], regulating metabolic rate and promoting cognitive function in childhood [thyroid-stimulating hormone (TSH)], responding to stress [adrenocorticotropic hormone (ACTH)], promoting linear growth in childhood [growth hormone (GH)], and stimulating lactation (prolactin). Posterior pituitary hormones regulate water balance [antidiuretic hormone (ADH)] and stimulate uterine contraction during childbirth and milk letdown following delivery (oxytocin). It is important to note that some of these hormones exert their effects on a single end-organ target (LH, FSH, TSH, and ACTH), while others act on multiple sites throughout the body (GH, prolactin, ADH, and oxytocin).

Anatomy

The pituitary is located at the base of the brain, immediately inferior to the hypothalamus, and is encased by the bony sella turcica. Although it plays a crucially important role, the pituitary is small with an approximate weight of 0.5−0.6 g. It is typically larger in women than men and reaches its largest size at the conclusion of pregnancy. During this period of enlargement, the pituitary is particularly susceptible to infarction if a woman experiences substantial blood loss during childbirth, leading to hypovolemic shock. This results in postpartum panhypopituitarism (pituitary hormone insufficiency) known as Sheehan's syndrome.

While the anterior and posterior pituitaries are located within close proximity of one another, the mechanisms responsible for hormone synthesis and release in the two lobes are fundamentally different. The posterior pituitary is an extension of the infundibulum, the narrow stalk-like structure that protrudes from the base of the hypothalamus toward the pituitary (Fig. 39.1). Because of this relationship, the posterior pituitary can be considered as an extension of the hypothalamus. Neurosecretory cells in the hypothalamus synthesize ADH and oxytocin, which are transported through the axons of these neurosecretory cells through the infundibulum to the posterior pituitary. Following appropriate hypothalamic stimuli, the posterior pituitary releases ADH and/or oxytocin into the surrounding capillary network for distribution throughout the body. Because the role of the posterior pituitary is limited to hormone release rather than synthesis, it is viewed as a reservoir for hormones synthesized upstream in the hypothalamus rather than a discrete endocrine organ.

By contrast, the anterior pituitary is not an extension of the hypothalamus but is a physically distinct entity (Fig. 39.1). The hypothalamus regulates anterior pituitary

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00039-9

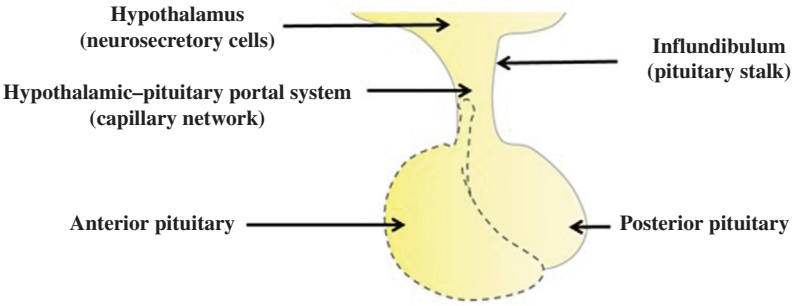


FIGURE 39.1 Hypothalamus and pituitary gland anatomy.

function by releasing regulatory hormones into the hypothalamus−pituitary portal system, a capillary network that connects the anterior pituitary and hypothalamus. These hypothalamic regulatory hormones either stimulate or repress hormone synthesis by specialized cells in the anterior pituitary. Approximately 50% of the anterior pituitary hormone-secreting cells are somatotrophs (GH-secreting cells), 10%−25% are lactotrophs (prolactin-secreting cells), 10% are gonadotrophs (FSH- and LH-secreting cells), 10%−20% are corticotrophs (ACTH-secreting cells), and 10% are thyrotrophs (TSH-secreting cells) [1]. In contrast to the posterior pituitary, the cells of the anterior pituitary are capable of hormone synthesis in addition to storage and release. The hypothalamus−pituitary portal system also permits retrograde flow, allowing anterior pituitary hormones to travel upstream to the hypothalamus and inhibit further release of hypothalamic regulatory hormones.

Physiology and regulation

The synthesis and release of most anterior pituitary hormones are stimulated by hypothalamic regulatory hormones that bind to specific receptors on the surface of hormone-producing anterior pituitary cells. With the exception of ACTH, anterior pituitary hormones are larger than their corresponding hypothalamic regulatory hormones (Table 39.1). Gonadotropin-releasing hormone (GnRH) binds to the GnRH receptors on gonadotrophs, stimulating the synthesis and release of LH and FSH. Similarly, thyrotropin-releasing hormone (TRH) stimulates thyrotrophs to produce TSH, corticotropin-releasing hormone (CRH) stimulates corticotrophs to produce ACTH, and growth hormone-releasing hormone (GHRH) stimulates somatotrophs to produce GH. Secreted anterior pituitary tropic hormone(s) travel through the vasculature to target endocrine tissue(s) and stimulate the production and release of endocrine secretory hormone(s). In order to prevent uncontrolled pituitary hormone synthesis, LH, FSH, TSH, ACTH, and GH are subject to negative feedback inhibition. The products synthesized by endocrine glands in response to these pituitary hormones bind to receptors in the hypothalamus and pituitary, preventing further release of the appropriate pituitary hormone. Feedback inhibition of LH, FSH, ACTH, and GH occurs primarily at the level of the hypothalamus, while inhibition of TSH occurs primarily at the level of the anterior pituitary (Fig. 39.2).

There are some notable exceptions and additions to the regulatory mechanism illustrated in Fig. 39.2. Prolactin is the only anterior pituitary hormone whose release is under a constant state of suppression. The hypothalamus produces dopamine that prevents prolactin release, leading to its alternate name of prolactin release inhibiting hormone (PRIH). By contrast, high concentrations of TRH observed in secondary hypothyroidism may stimulate prolactin release. GH regulation by the hypothalamus is also unique in that the hypothalamus primarily stimulates GH release through GHRH, and can also inhibit GH release through somatostatin [somatotropin release inhibiting hormone (SRIH)]. Lastly, while ACTH production is predominantly stimulated by hypothalamic CRH, it can also be stimulated by certain inflammatory cytokines or very high concentrations of ADH.

The mechanisms controlling the release of the posterior pituitary hormones ADH and oxytocin are very different from the mechanism shown in Fig. 39.2, and will be described later in this chapter.

Anterior pituitary hormone pathophysiology and lab findings

Disorders caused by atypical pituitary function can be divided into two categories: those marked by excessive pituitary hormone production (hyperfunction) and those marked by deficient pituitary hormone production (hypofunction). A summary of anterior pituitary disorders is provided in Table 39.2.

Hyperfunction

With the exception of hyperprolactinemia, which has many nonneoplastic causes, anterior pituitary adenomas are the most common cause of anterior pituitary hyperfunction.

TABLE 39.1 Anterior pituitary hormones and their source, composition, and associated hypothalamic regulatory hormone.

Hormone	Abbreviation (alternative name)	Source	Subunits	Molecular weight (amino acids)	Hypothalamic regulatory hormone (amino acids)
Adrenocorticotropic hormone	ACTH (corticotropin)	Corticotroph	1	4.5 kDa (39)	CRH (41)
Thyroid-stimulating hormone	TSH (thyrotropin)	Thyrotroph	2	28 kDa (alpha: 92; beta: 118)	TRH (3)
Growth hormone	GH (somatotropin)	Somatotroph	1	22 kDa (191) 20 kDa (176)	GHRH (44) SRIH (14)
Luteinizing hormone	LH (gonadotropin)	Gonadotroph	2	32 kDa (alpha: 92; beta: 121)	GnRH (10)
Follicle-stimulating hormone	FSH (gonadotropin)	Gonadotroph	2	30 kDa (alpha: 92; beta: 111)	GnRH (10)
Prolactin	PRL (lactotropin)	Lactotroph (mammotroph)	1	22 kDa (198)	PRIH

CRH, Corticotropin-releasing hormone; *GHRH*, growth hormone-releasing hormone; *GnRH*, gonadotropin-releasing hormone; *PRIH*, prolactin release inhibiting hormone; *SRIH*, somatotropin release inhibiting hormone; *TRH*, thyrotropin-releasing hormone.

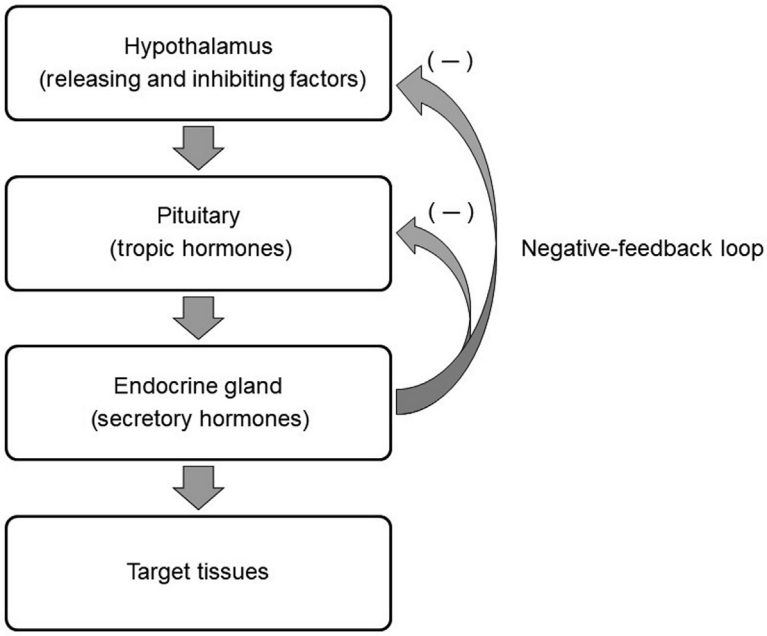


FIGURE 39.2 Functional interrelationship of the hypothalamus, pituitary, and peripheral endocrine glands.

TABLE 39.2 Disorders of underproduction and overproduction of anterior pituitary hormones.

Anterior pituitary hormone	Consequences of hormone excess	Consequences of hormone deficiency
ACTH	Cushing's disease	Cortisol deficiency
TSH	Central hyperthyroidism	Central hypothyroidism
GH	Children: gigantism Adults: acromegaly	Children: short stature Adults: adult growth hormone deficiency
LH and FSH	Generally asymptomatic[a]	Hypogonadism
Prolactin	Galactorrhea and hypogonadism	Inadequate lactation

LH, Luteinizing hormone; *FSH*, follicle-stimulating hormone; *TSH*, thyroid-stimulating hormone; *ACTH*, adrenocorticotropic hormone; *GH*, growth hormone.
[a]*Pituitary gonadotropinomas typically secrete free alpha subunit rather than intact LH and FSH.*
Source: Adapted from W. Winter, N.S. Hardt, N.S. Harris, Disorders of the anterior and posterior pituitary. *Contemporary Practices in Clinical Chemistry*, third edn., 2016.

Adenomas are discovered either during the evaluation of symptoms caused by excessive anterior pituitary hormone production or following complaints of vision changes, persistent headache, impaired sense of smell, or other complications related to the compression of neighboring tissues by the growing tumor, a phenomenon known as "mass effect." Adenomas less than 1 cm are known as microadenomas, and these tumors often secrete anterior pituitary hormones. Microadenomas are discovered relatively early, because the excessive pituitary hormone production results in altered endocrine function, prompting patients to seek medical care. By contrast, adenomas 1 cm or greater are known as macroadenomas. These tumors often do not secrete anterior pituitary hormones, and as a result, patients with a macroadenoma may not experience clearly identifiable symptoms consistent with altered endocrine function. This may cause delay in diagnosis until the tumor reaches a sufficient size to interfere with the function of neighboring tissue. Prolactinomas are the most common anterior pituitary adenoma, followed in frequency by nonsecretory gonadotropinomas (nonsecretory is a bit of a misnomer as these tumors may produce free alpha subunit but not the biologically active intact alpha/beta heterodimer), corticotropinomas (ACTH), somatotropinomas (GH), and thyrotropinomas (TSH, exceedingly rare) [2]. Occasionally, hormone-secreting adenomas may produce and secrete multiple anterior pituitary hormones (e.g., GH and prolactin cosecretion), but production of a single anterior pituitary hormone is more common, as these tumors arise from a single specialized cell that maintains many of its original characteristics.

Hypofunction

Disorders of anterior pituitary hypofunction can be caused by hypothalamic disease, pituitary disease, or disruption of the hypothalamus–pituitary portal system (not true for prolactin, which is increased in cases of hypothalamus–pituitary portal system disruption). In contrast to the short list of conditions responsible for anterior pituitary hyperfunction, there are many possible causes of pituitary hypofunction, including infection, inflammation, trauma, surgical or radiological treatment of a previously diagnosed tumor, congenital malformations, or infarction (Sheehan's syndrome and pituitary apoplexy). As these mechanisms are not specific to one unique anterior pituitary cell type, they typically result in deficiencies in many (and often all) anterior pituitary hormones, a phenomenon known as panhypopituitarism. Single hormone deficiency may result from genetic defects in anterior pituitary hormones themselves, hormone receptors, or other hormone-specific signal transduction pathways that disrupt the function of only one anterior pituitary hormone while leaving the remaining anterior pituitary hormone regulatory pathways intact.

Adrenocorticotropic hormone

ACTH, also referred to as adrenocorticotropin and corticotropin, plays a key role in stimulating adrenal glucocorticoid (e.g., cortisol) production, which is centrally regulated by the hypothalamus–pituitary–adrenocortical axis (Fig. 39.3). The hypothalamus produces corticotropin-releasing hormone (CRH; 41 amino acid polypeptide) that travels to the anterior pituitary via the hypothalamus–pituitary portal system. In the anterior pituitary, CRH binds to receptors on corticotropic cells (i.e., corticotrophs or corticotropes) and stimulates ACTH production. ACTH is a peptide hormone product of the precursor protein proopiomelanocortin (POMC). POMC is cleaved to produce ACTH as well as additional fragments, including melanocyte-stimulating hormones (MSHs), lipotropins, and β-endorphin (Fig. 39.4). Following release into circulation, ACTH binds to receptors on cells in the adrenal cortex to stimulate cortisol production.

An important biological variable for many endocrine hormones, including ACTH and cortisol, is the cyclic production pattern over a 24-hour period, which is referred to as circadian rhythm or diurnal variation. Cortisol exhibits diurnal release where peak concentration is typically found in the early morning and nadir concentration in the evening. Similarly, plasma ACTH concentration is typically highest in the early morning. This cyclic pattern of release is regulated by neurotransmitters in the

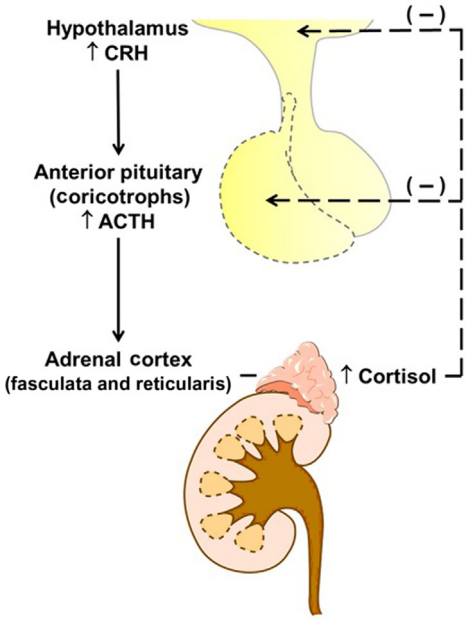


FIGURE 39.3 Hypothalamus–pituitary–adrenal axis. Adrenal production of cortisol in the zona fasciculata is stimulated by pituitary adrenocorticotropic hormone, which, in turn, is regulated by hypothalamic corticotropin-releasing hormone. Corticotropin-releasing hormone is predominantly responsible for stimulating adrenocorticotropic hormone production, while cortisol provides negative feedback primarily at the level of the hypothalamus and to a lesser extent at the pituitary.

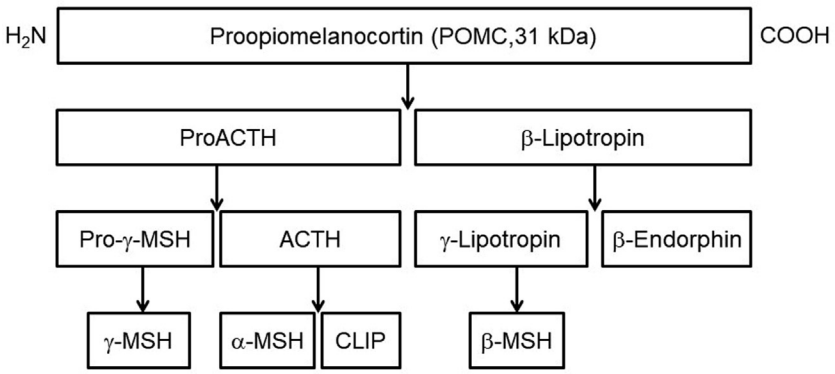


FIGURE 39.4 Proopiomelanocortin and its precursor relationship to derivative hormones (adrenocorticotropic hormone, lipotropins, melanocyte-stimulating hormone, and endorphins) produced by proteolytic degradation.

hypothalamus and influenced by each individual's biological clock. For example, a night-shift worker's sleep and eating patterns can alter the typical timing pattern of peak and nadir cortisol concentrations. It is important to consider preanalytical biological variation in the interpretation of patient ACTH and cortisol results.

Pituitary disease resulting in overproduction of ACTH can cause dysregulation in the downstream production of cortisol from the adrenal cortex. Cortisol is a glucocorticoid that plays a critical role in glucose metabolism and in the body's response to stress. Signs, symptoms, and sequelae associated with hypercortisolism (i.e., cortisol excess) may include central obesity, hyperglycemia, hypertension, hirsutism, muscle weakness, and others, which is commonly referred to as Cushing's syndrome. There are numerous potential etiologies for hypercortisolism, which are categorized as ACTH-dependent or ACTH-independent. In primary adrenal disease, hypercortisolism is ACTH-independent and may result from overproduction of cortisol by the adrenal gland(s) (e.g., cortisol-secreting adrenocortical adenoma or carcinoma, and adrenal hyperplasia) or from exogenous glucocorticoid (e.g., prednisone) administration used as treatment in a variety of medical conditions. The hypothalamus−pituitary−adrenal axis negative feedback regulatory loop is maintained in primary adrenal disease; accordingly, the excess cortisol causes suppression of CRH and ACTH production. In Cushing's disease, hypercortisolism is due to a functioning secretory pituitary tumor (e.g., adenoma) and accounts for 75%−80% of ACTH-dependent hypercortisolism [3]. Overproduction of ACTH by a functioning adrenal adenoma causes ACTH excess and continuous stimulation of the adrenal glands, leading to hyperplasia and overproduction of cortisol. ACTH-dependent hypercortisolism can also occur from a nonpituitary ectopic source, such as bronchial carcinoid or small cell lung cancer. Ectopic ACTH production accounts for 15%−20% of ACTH-dependent hypercortisolism and is associated with suppression of CRH production and pituitary-derived ACTH [3]. Additional information on hypercortisolism can be found in Chapter 41, Disorders of the adrenal cortex and medulla, of this book.

Pituitary disease resulting in ACTH deficiency can also cause endocrine dysregulation with inadequate production of cortisol from the adrenal cortex. Signs, symptoms, and sequelae associated with hypocortisolism (i.e., glucocorticoid deficiency) may include weakness, malaise, weight loss, hypoglycemia, and others. There are numerous potential etiologies for glucocorticoid deficiency, which are stratified based on a mechanism of dysfunction that may occur in the adrenal cortex, anterior pituitary, or hypothalamic gland. Addison's disease, also referred to as primary glucocorticoid deficiency or primary adrenal insufficiency (with concomitant mineralocorticoid deficiency), is a disorder of the adrenal glands, resulting in the insufficient production of cortisol that can be life-threatening. Patients with Addison's disease may appear hyperpigmented (e.g., skin appears bronze color) on presentation due to increased production of MSH in concert with ACTH (Fig. 39.4). Secondary glucocorticoid deficiency refers to a form of hypopituitarism that causes insufficient production of ACTH; consequently, there is diminished stimulation of the adrenal cortex, leading to low circulating cortisol concentrations. Tertiary glucocorticoid deficiency occurs when cortisol deficiency is due to a hypothalamic disorder that causes inadequate CRH production and downstream hypopituitarism with insufficient ACTH production.

Laboratory testing is the key to diagnosis of hypercortisolism or hypocortisolism and identifying the etiology of hormone dysregulation. Nonspecific signs and symptoms of cortisol abnormalities as well as various etiologies of potential dysfunction in the hypothalamus−pituitary−adrenocortical axis collectively contribute to challenging differential diagnoses. It is important to determine the etiology of hypercortisolism or hypocortisolism for appropriate patient treatment and management.

The first step in the evaluation of suspected Cushing's syndrome is to confirm hypercortisolemia. Tests for hypercortisolemia include 24-hour urinary free cortisol, overnight low-dose dexamethasone suppression testing, and midnight salivary cortisol testing. These tests are

TABLE 39.3 Summary of expected results in the evaluation of hypercortisolism or hypocortisolism.

Pathophysiology	24-h urinary free cortisol	Plasma ACTH
Primary adrenal disease (primary hypercortisolism)	↑	↓
Cushing's disease	↑	↑
Ectopic ACTH	↑	↑
Addison's disease (primary glucocorticoid deficiency)	↓	↑
Secondary glucocorticoid deficiency	↓	↓
Tertiary glucocorticoid deficiency	↓	↓

ACTH, Adrenocorticotropic hormone.

described in detail in Chapter 41, Disorders of the adrenal cortex and medulla, of this book. The 24-hour urinary free cortisol test is generally preferred to assess potential hypercortisolism, whereas a random serum cortisol test is of little clinical value, considering the diurnal variation in cortisol production. An elevated 24-hour urinary free cortisol result supports the presence of hypercortisolemia, and subsequent testing aims to localize the source of cortisol excess. Measurement of plasma ACTH is the key, as plasma ACTH concentrations will be suppressed if the source of hypercortisolemia is the adrenal gland itself. Conversely, plasma ACTH will be within or elevated above reference interval if there is an ACTH-producing lesion in either the pituitary gland or an ectopic location (Table 39.3). In such cases, dexamethasone suppression testing can help identify the location of excess ACTH production.

Dexamethasone suppression test

Dexamethasone suppression testing protocols can vary, and, in general, this test involves repeated administration of low- and/or high-dose dexamethasone (e.g., 0.5 or 2 mg every 6 hours for 2 days) with serial collection of serum and plasma samples for cortisol and ACTH measurements, respectively. Patients with hypercortisolemia due to constitutive production of ACTH from an ectopic source show minimal suppression of ACTH production following high-dose dexamethasone suppression testing (HDDST; Fig. 39.5A). Conversely, patients with Cushing's disease still retain some degree of negative feedback in the hypothalamus–pituitary–adrenocortical axis and show suppression of ACTH following HDDST (Fig. 39.5B). Radiologic imaging studies of the pituitary can be helpful; however, ACTH-secreting pituitary tumors are often too small to visualize. Bilateral inferior pituitary sinus sampling (BIPSS) can also help determine the source of excess ACTH production. The inferior petrosal sinuses receive drainage from the pituitary gland without admixture with blood from other sources. Catheters are advanced bilaterally to the inferior petrosal sinuses in order to obtain samples localized at the pituitary, along with simultaneous peripheral venous blood collection. The collected pituitary and venous samples are tested for ACTH, and the pituitary-to-peripheral ACTH ratios are calculated. An elevated pituitary-to-peripheral ACTH ratio (e.g., ratio >3) is indicative of Cushing's disease, whereas a low ratio (e.g., ratio <2) supports an adrenal or ectopic source of ACTH. BIPSS may not be available at all centers, because this is an invasive procedure that should be performed by expert vascular radiologists at experienced centers. Expected hormone profiles in healthy individuals or patients with primary adrenal disease are shown for comparison (Fig. 39.5C and D); however, dexamethasone suppression testing is typically unnecessary in these cases.

Laboratory testing for suspected glucocorticoid deficiency begins with confirmation of hypocortisolism followed by dynamic testing to identify the etiology of dysfunction in the hypothalamus–pituitary–adrenocortical axis. Dynamic testing under controlled conditions can help determine the etiology of hormone dysregulation and may include one or more of the following: (1) ACTH (e.g., cosyntropin and synacthen) stimulation test; (2) insulin tolerance test (ITT); (3) metyrapone test; and (4) CRH stimulation test.

Adrenocorticotropic hormone stimulation test

Cosyntropin is an ACTH derivative that includes the first 24 amino acids of ACTH and maintains ACTH activity. The ACTH stimulation test generally involves intravenous (IV) or intramuscular administration of cosyntropin (e.g., 250 μg), with collection of serum and measurement of cortisol at baseline as well as 30–120-minute poststimulation. Under normal physiological conditions, the adrenal gland responds to cosyntropin administration with increased production of cortisol to a peak concentration of $\geq 18–20$ μg/dL (Fig. 39.6) [4]. In primary adrenal insufficiency (e.g., Addison's disease), there is diminished cortisol production in response to cosyntropin, because the adrenal cortex is dysfunctional, damaged, or destroyed. Finding an elevated basal ACTH concentration

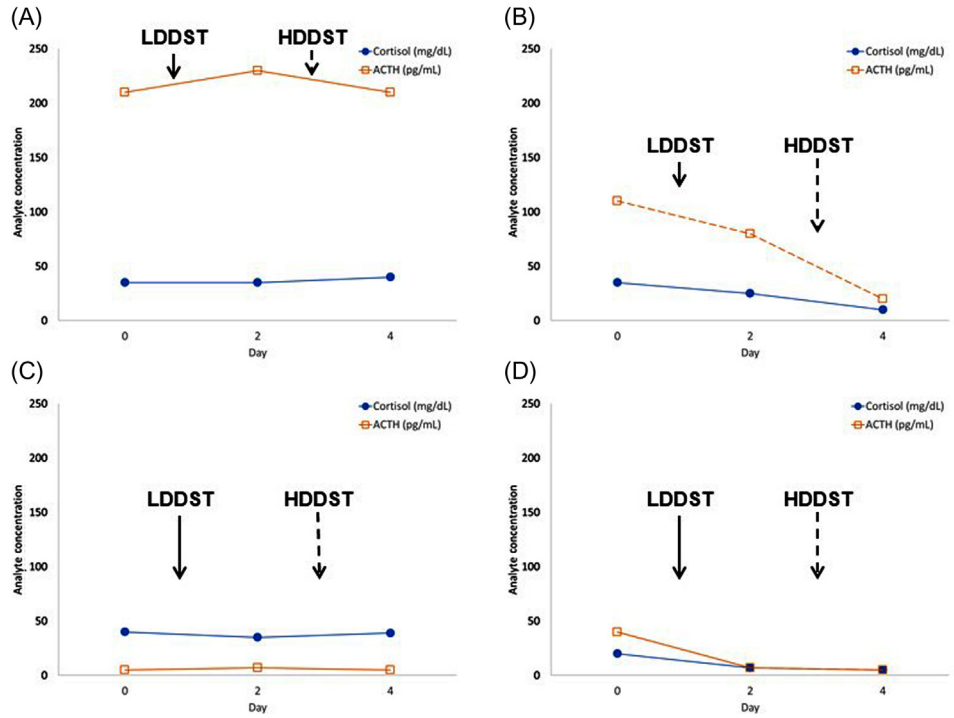


FIGURE 39.5 Expected hormone responses to low-dose and high-dose dexamethasone suppression tests. (A) Expected adrenocorticotropic hormone and cortisol results for ectopic adrenocorticotropic hormone syndrome. High plasma concentrations of adrenocorticotropic hormone and cortisol are unresponsive to low-dose dexamethasone suppression testing and high-dose dexamethasone suppression testing. (B) Expected adrenocorticotropic hormone and cortisol results for Cushing's disease. Suppression of the elevated adrenocorticotropic hormone and cortisol are incomplete with low-dose dexamethasone suppression testing, but >50% suppression with high-dose dexamethasone suppression testing. (C) Expected adrenocorticotropic hormone and cortisol results for a cortisol-secreting adrenal tumor. Autonomous cortisol production by the adrenal tumor causes adrenocorticotropic hormone suppression, even before administration of dexamethasone. (D) Expected adrenocorticotropic hormone and cortisol results with normal hypothalamus–pituitary–adrenal function (no cortisol excess). Plasma adrenocorticotropic hormone and cortisol results are within reference intervals at baseline and become markedly suppressed by low-dose dexamethasone suppression testing.

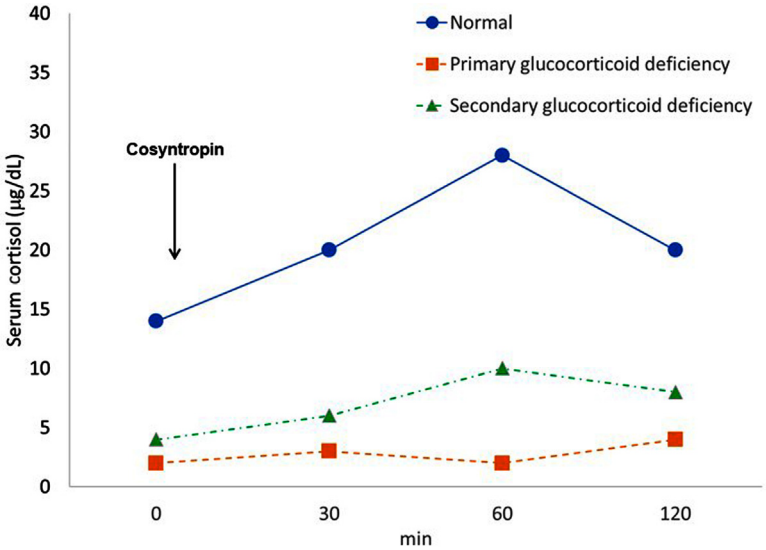


FIGURE 39.6 Expected cortisol responses to the cosyntropin (adrenocorticotropic hormone) stimulation test. The squares represent expected cortisol response in primary adrenal insufficiency showing a low baseline cortisol concentration that is unresponsive to the adrenocorticotropic hormone stimulation test. The expected cortisol response with secondary adrenal insufficiency (triangles) often shows a mild increase in serum cortisol with adrenocorticotropic hormone stimulation; however, the cortisol response is blunted due to adrenal atrophy that can occur with adrenocorticotropic hormone insufficiency. The circles represent expected cortisol response under normal physiological conditions. Serum cortisol is within reference interval at baseline and rises >18 μg/dL after adrenocorticotropic hormone stimulation.

is also consistent with primary dysfunction in cortisol production. In secondary or tertiary glucocorticoid deficiency, an increase in cortisol production following cosyntropin stimulation is expected; however, if ACTH deficiency is severe and prolonged, it may have induced adrenal atrophy, resulting in a blunted or diminished cortisol response to cosyntropin (Fig. 39.6) [5].

Insulin tolerance test

The ITT can be used to assess glucocorticoid deficiency, because hypoglycemia is a strong stressor that stimulates rapid activation of the hypothalamus–pituitary–adrenocortical axis. The ITT test generally involves IV administration of insulin (e.g., 0.1 U/kg) following an overnight fast, with collection of samples at baseline, 30, 60, 90, and 120 minutes (times may vary) for cortisol, glucose, and ACTH testing. Hypoglycemia is achieved when the nadir glucose concentration is reduced >50% from baseline or <45 mg/dL [5]. Under normal physiological conditions, ITT should induce a cortisol response of ≥20 μg/dL. In primary glucocorticoid deficiency, the cortisol response is deficient, and an elevated ACTH value is observed. Conversely, in secondary or tertiary glucocorticoid deficiency, both cortisol and ACTH responses are deficient. The ITT must be performed with caution under experienced medical supervision because of the risks of hypoglycemia (e.g., obtundation, seizures, and death).

Metyrapone test

Metyrapone is an inhibitor of the 11β-hydroxylase enzyme that normally catalyzes the conversion of 11-deoxycortisol to cortisol during steroid synthesis, thereby decreasing the negative feedback of cortisol in the hypothalamus–pituitary–adrenocortical axis. The overnight metyrapone test involves oral administration of metyrapone (e.g., 30 mg/kg) at midnight with collection of plasma or serum samples the following morning for measurement of cortisol and 11-deoxycortisol. Criteria for defining a normal metyrapone test vary. In general, a normal physiological response to metyrapone inhibition results in accumulation of the precursor 11-deoxycortisol, reduced cortisol, and increased ACTH production. Abnormal findings further support adrenal insufficiency. The metyrapone test is rarely performed and not recommended in suspected primary glucocorticoid deficiency because of serious risk of precipitating an adrenal crisis (Addisonian crisis).

Corticotropin-releasing hormone stimulation test

The CRH stimulation test can be used to help differentiate between secondary (pituitary) and tertiary (hypothalamic) glucocorticoid deficiencies. The CRH test involves administration of an IV bolus of ovine CRH (e.g., 1 μg/kg) followed by serial collection of samples (e.g., baseline, +5, +10, +15, +30, +45, +60, +90, and +120 minutes) for ACTH testing. Under normal physiological conditions, there may be doubling of ACTH concentration at 10–15 poststimulation compared with baseline value, or stimulation of ACTH values of 20–100 pg/mL. However, individual responses to exogenous CRH are highly variable, and thresholds for normal responses are not well-defined, which limit the clinical usefulness of this test [5]. A normal ACTH response to CRH stimulation indicates tertiary glucocorticoid deficiency (i.e., CRH deficiency) and some form of hypothalamic disturbance or dysfunction in the hypothalamus–pituitary portal system. In secondary glucocorticoid deficiency (i.e., ACTH deficiency), there is a deficient ACTH response to CRH stimulation, supporting a pituitary disorder.

Thyroid-stimulating hormone

TSH, produced by thyrotrophs of the anterior pituitary, stimulates thyroid follicular cells to produce the thyroid hormones T4 (thyroxine) and T3 (triiodothyronine) by binding to transmembrane receptors expressed on the follicular cell surface. T4 and T3 play important roles in fetal and early childhood neurological development, promote linear growth in children, and regulate the metabolic rate throughout life. TSH is a member of the glycoprotein hormone family that also includes LH, FSH, and human chorionic gonadotropin (hCG). All four of these hormones are heterodimers, consisting of an identical alpha subunit and a unique beta subunit that confers biological specificity (Fig. 39.7). However, the beta subunits show substantial sequence homology, and cross-reactivity with other glycoprotein hormone receptors has been observed at high concentrations. For example, high concentrations of hCG consistent with early pregnancy can activate the TSH receptor, stimulating T4 and T3 production.

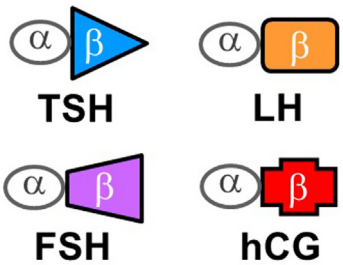


FIGURE 39.7 Glycoprotein hormone family members include thyroid-stimulating hormone, follicle-stimulating hormone, luteinizing hormone, and human chorionic gonadotropin. The alpha subunit is identical for all four hormones, and the beta subunit confers biological specificity. The beta subunits are highly conserved, and cross-reactivity with other glycoprotein hormone receptors may occur at high hormone concentrations.

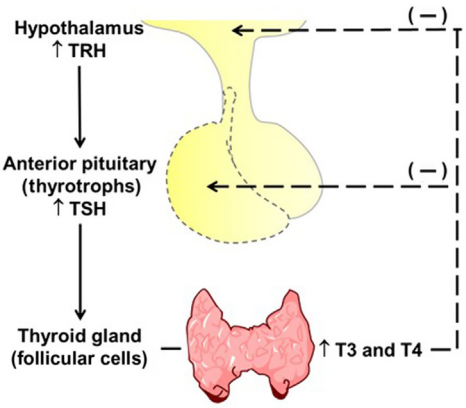


FIGURE 39.8 Hypothalamus−pituitary−thyroid axis. Thyroid-stimulating hormone production by the anterior pituitary is stimulated by hypothalamic thyrotropin-releasing hormone and suppressed by free thyroxine (fT4), which feeds back primarily at the level of the pituitary and to a lesser extent at the hypothalamus.

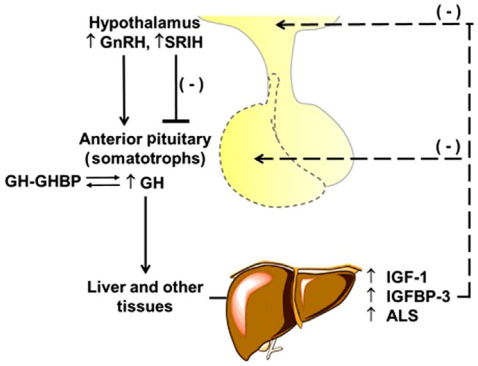


FIGURE 39.9 Growth hormone regulatory axis. Hypothalamic growth hormone-releasing hormone promotes growth hormone release by the pituitary, while hypothalamic somatotropin release inhibiting hormone suppresses growth hormone release. Growth hormone acts throughout the body and also stimulates hepatic insulin-like growth factor 1, insulin-like growth factor binding protein 3, and acid-labile subunit production, which provide negative feedback primarily on the hypothalamus and to a lesser extent on the pituitary.

TSH secretion is controlled by both positive stimulus in the form of TRH produced by the hypothalamus and negative feedback inhibition by the biologically active, nonprotein bound fraction of T4 (free T4 or fT4) (Fig. 39.8). The TSH regulatory axis is unique in that free T4 feeds back primarily at the level of the pituitary, while endocrine products that inhibit the production of other anterior pituitary hormones act primarily at the level of the hypothalamus. The anterior pituitary is exquisitely sensitive to plasma free T4 concentrations, and thyrotrophs rapidly increase or decrease TSH production to maintain free T4 within a narrow physiologically appropriate range. Provocative testing of the hypothalamus−pituitary−thyroid axis is not required, because even mild thyroid dysfunction causes TSH concentrations to fall outside of the expected reference interval. The development of third generation TSH immunoassays with a lower limit of detection of 0.01 μIU/mL has made it possible to identify states of hyperthyroidism, hypothyroidism, and euthyroidism with a single screening test, although many providers choose to measure TSH and free T4 simultaneously. It is important to note that TSH, like the other anterior pituitary hormones, is secreted in a pulsatile pattern and has a short half-life of 1−2 hours. Despite this, TSH and free T4 can be measured any time of day, as they do not show substantial diurnal variation. Intraindividual variation in TSH is approximately 20% over time intervals of 1 week to 1 year.

Thyroid hormone synthesis, transport, biological activity, and analytical measurement methods are discussed in detail in Chapter 40, Laboratory evaluation of thyroid function.

Growth hormone

In contrast to the other anterior pituitary hormones, GH is not limited to a single target endocrine gland but rather acts on many different sites throughout the body that express GH receptors. Most important among these are the liver and sites of bone growth during childhood. In the liver, GH stimulates the production of insulin-like growth factor 1 (IGF-1), insulin-like growth factor binding protein-3 (IGFBP-3), and the acid-labile subunit. These three proteins form a complex that provides a reserve of circulating IGF-1 that maintains stable concentrations of serum free IGF-1 (not bound to GH-binding protein). It was initially thought that the liver was the sole source of IGF-1 production, but it is currently accepted that, while most plasma IGF-1 is hepatic in origin, many cell types produce IGF-1 that has autocrine or paracrine functions. GH production and release by the anterior pituitary are largely stimulated by hypothalamic GHRH, but hypothalamic SRIH does provide a modest inhibitory signal. IGF-1 also prevents GH secretion through feedback inhibition, primarily at the level of the hypothalamus (Fig. 39.9).

GH directly elicits some of its biological effects, while others are mediated by IGF-1. GH directly raises plasma glucose by inducing insulin resistance and also increases free fatty acid plasma concentrations. In contrast, IGF-1 is primarily responsible for promoting linear growth in children, but it is important to note that locally produced IGF-1 has a greater effect on growth than circulating hepatic IGF-1. Adding further complexity is the fact that two distinct forms of GH are secreted by the anterior pituitary: a 191 amino acid (22 kDa) complete form and a 176 amino acid (20 kDa) form produced by alternative splicing in which a portion of exon 3 that encodes amino acids 32−46 is removed. However, the two forms demonstrate equivalent growth-promoting and lipolytic properties, and the only substantial physiological difference between the two is the diminished lactogenic activity

exhibited by the 20-kDa form. The larger 22-kDa form constitutes 80%−85% of GH found in circulation.

GH deficiency is generally not a concern in the adult population but should be considered in children with short stature and slow growth velocity once other causes including Turner's syndrome, hypothyroidism, malnutrition, and chronic illness have been excluded [6]. Because GH is secreted in short pulses that typically occur at night, measurement of GH in a randomly collected plasma or serum specimen is unlikely to be informative. For this reason, initial screening for GH deficiency is typically performed by measuring IGF-1 as a surrogate marker for GH activity that demonstrates more constant plasma concentrations throughout the day. Plasma IGFBP-3 also reflects GH activity but rarely provides additional information beyond measuring IGF-1 alone. If IGF-1 is within the age/gender-appropriate reference interval, GH deficiency can be ruled out. However, a low IGF-1 result does not confirm GH deficiency, as GH is not the sole determinant of IGF-1 concentration. Low IGF-1 can also be observed in patients with hypothyroidism, malnutrition, malabsorption, renal failure, chronic inflammation, and other disease states. While most providers choose to screen for GH deficiency by measuring IGF-1, screening can also be accomplished by performing provocative testing that assesses the patient's ability to secrete GH in response to stimulus. Using this approach, GH is measured in a single specimen collected after 20 minutes of vigorous exercise or 60−90 minutes after administration of a compound known to stimulate GH release, including L-dopa (non-United States), clonidine, arginine, or glucagon. Depending on the assay used, the appropriate physiological response is a stimulated GH concentration between 7 and 10 ng/mL [7].

Following an abnormal screening test, the appropriate next step is a formal stimulation test. These are conceptually similar to the screening tests but require the collection of multiple specimens at regular intervals following pharmacologic stimulus. In addition to L-dopa, clonidine, arginine, and glucagon, insulin may also be used for definitive testing. However, as described previously, other pharmacologic options are frequently preferred over insulin. Recommended doses and times of collection for the formal stimulation tests are listed in Table 39.4. In practice, it is common to perform two sequential stimulation tests using different pharmacologic compounds (e.g., clonidine followed by arginine). Like the screening tests, a single GH concentration between 7 and 10 ng/mL at any time point following administration of either stimulatory compound is considered an appropriate response. Stimulated GH values of <7 ng/mL in children and <3−5 ng/mL in adults are consistent with GH deficiency. In order to reduce the occurrence of falsely low GH stimulation test results in peripubertal children without GH deficiency, pretreatment with sex steroids may be performed (Table 39.5). However, this practice is controversial, and several studies have documented no statistically significant difference in the diagnostic yield

TABLE 39.5 Sex steroid priming prior to growth hormone testing in peripubertal children.

Agent	Dose
Ethinyl estradiol	50−100 μg/day × 3 days
Premarin	5 mg/day × 3 days
Diethylstilbestrol	5 mg/day × 3 days
Testosterone	100 mg/day × 3 days

Source: Adapted from W. Winter, N.S. Hardt, N.S. Harris, Disorders of the anterior and posterior pituitary. *Contemporary Practices in Clinical Chemistry*, third edn., 2016.

TABLE 39.4 Amino acid and pharmacologic stimuli used in growth hormone testing.

GH stimulus	Dose	Sampling time (min)[a]
Arginine HCl	0.5 g/kg and max 30 g	0, ±15, 30, 60, 90, and 120
Glucagon	0.1 mg/kg and max 1 mg	0, 30, 60, 90, 120, 150, 180, and ±240
L-Dopa	10 mg/kg and max 500 mg	0, 40, 60, 90, and 120
Clonidine	0.15 mg/m^2	0, 30, 60, and 90
Insulin	0.10 U regular insulin/kg	0, 30, 60, and 90[b]

GH, Growth hormone.
[a]*Expert opinion varies on the optimal interval for poststimulus growth hormone measurement.*
[b]*Evidence of adequate pituitary stimulation consists of hypoglycemic symptoms (sweating and tremor) or plasma glucose concentrations of <45 mg/dL.*
Source: Adapted from W. Winter, N.S. Hardt, N.S. Harris, Disorders of the anterior and posterior pituitary. *Contemporary Practices in Clinical Chemistry*, third edn., 2016.

of GH stimulation tests between patients who undergo sex steroid priming and those who do not [8].

In very rare cases, individuals with clinical signs and symptoms of GH deficiency may have elevated GH values and low IGF-1 values in a random plasma specimen. Once more common causes of GH insufficiency, such as malnutrition, renal failure, and hepatic disease have been excluded, GH resistance should be considered. Mutations in the GH receptor or downstream signaling proteins have been identified in patients with GH resistance.

GH excess that develops in adulthood is known as acromegaly and is characterized by physical changes that include enlargement of the hands, feet, heart and/or liver, coarsened facial features, metabolic changes that include insulin resistance/diabetes mellitus, increased sweating, myopathy, and premature cardiovascular disease [9]. Ninety-five percent of acromegaly cases are caused by anterior pituitary GH-secreting adenomas. Occasionally, these GH-secreting adenomas also secrete prolactin, resulting in galactorrhea and menstrual irregularities in women and decreased libido and breast enlargement in men, in addition to the acromegaly symptoms listed above. Five percent of acromegaly cases are caused by a hypothalamic GHRH-secreting tumor, and the remainder are due to GH- or GHRH-secreting tumors located outside of the hypothalamus and pituitary. GH excess that develops in childhood is known as gigantism and is marked by extreme tall stature due to excessive GH activity prior to fusion of the growth plates. In addition to tall stature, individuals with gigantism also exhibit the constellation of features observed in acromegaly.

Some cases of acromegaly may be recognized on the basis of physical features alone, but laboratory testing is frequently required to make a formal diagnosis, particularly in relatively mild cases with subtle physical changes. Most individuals with active acromegaly will have random GH concentrations above the upper limit of the reference interval, but a random GH measurement within the reference interval does not exclude disease. As in the workup of GH deficiency, plasma IGF-1 is a useful screening test, as random IGF-1 values often reflect disease severity more accurately than random GH measurements. If GH excess is suspected on the basis of physical observation or elevated plasma GH and/or IGF-1 measurements, provocative testing in the form of the glucose suppression test may be performed. Briefly, individuals undergoing the glucose suppression test fast overnight, remain at bed rest, and submit baseline plasma specimens for GH and glucose measurement prior to consuming 75-g glucose. Subsequent plasma specimens are collected 30, 60, 90, and 120 minutes later. As glucose suppresses GH release, plasma GH concentrations are expected to fall below the assay's lower limit of quantitation (typically 1 or 2 μg/L) in individuals without acromegaly. In patients with acromegaly, plasma GH is not suppressed and occasionally shows a paradoxical increase.

Most commercially available GH measurement methods are two site "sandwich" immunometric assays, but it is important to note that these immunoassays are poorly standardized due to the use of different antibodies and reference calibrator materials. Some assays selectively recognize the 22-kDa form, while others demonstrate significant cross-reactivity with both the 22- and 20-kDa forms, and yet others detect only the 20-kDa form. As a result, differences of 15%−20% may be observed when testing a single patient sample on multiple analytical platforms, which has important implications for GH stimulation tests that are interpreted in the context of a single pan-manufacturer cutoff. Given the variability between analytical platforms, the use of a pan-manufacturer cutoff is inappropriate and may lead to misinterpretation of stimulation test results. Unfortunately, manufacturer-specific cutoffs are not available, but significant efforts are aimed at harmonizing GH assays to eliminate this problem.

Recently, LC-MS/MS methods that provide independent measurements of both the 22- and 20-kDa GH forms have been developed for the detection of GH doping to improve athletic performance. These LC-MS/MS methods consist of an initial antibody-based enrichment of all GH forms, followed by proteolytic digestion and quantification of specific peptides associated with each proteoform. Natural GH production typically results in a ratio of ~12:1 (22−20 kDa), but exogenous supplementation with the 22-kDa form suppresses pituitary secretion, resulting in an elevated ratio.

Like GH assays, commercially available immunoassays for IGF-1 and IGFBP-3, as well as LC-MS/MS IGF-1 methods, are also poorly standardized. However, manufacturer-specific reference intervals are available for both pediatric and adult patients, which facilitates result interpretation.

Follicle-stimulating hormone and luteinizing hormone

FSH and LH are glycoprotein hormones produced by gonadotrophs of the anterior pituitary that are composed of an identical alpha subunit (also shared by TSH and hCG) and a unique beta subunit that confers biological specificity (Fig. 39.7). FSH is responsible for stimulating the growth and maturation of ovarian follicles, promoting ovarian estrogen synthesis and stimulating testicular spermatogenesis. LH is responsible for triggering ovulation and forming the corpus luteum at the midpoint of the menstrual cycle and stimulating the development and function of testicular Leydig cells that produce testosterone. Collectively, FSH and LH control gonadal function in both females and males. In states of reproductive health, plasma FSH and LH concentrations rise and fall in a tightly regulated pattern in response to positive and negative stimuli from

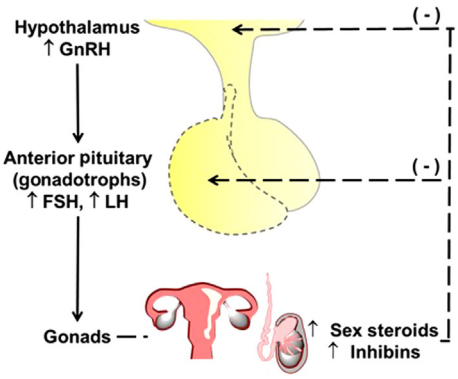


FIGURE 39.10 Hypothalamus–pituitary–gonadal axis. Follicle-stimulating hormone and luteinizing hormone production by the anterior pituitary is stimulated by hypothalamic gonadotropin-releasing hormone and suppressed by sex steroids and inhibins.

hypothalamic and gonadal hormone products. Alterations in this signaling pathway result in reproductive dysfunction, and hormone measurement in the clinical laboratory is often essential to identify the underlying pathology and select an appropriate course of treatment.

Similar to the other anterior pituitary hormones (with the exception of prolactin), FSH and LH release is positively stimulated by the hypothalamus (Fig. 39.10). However, this regulatory mechanism is unique in that a single hypothalamic releasing hormone (GnRH) promotes the synthesis and release of two different pituitary hormones (FSH and LH) from a single cell type. Furthermore, although FSH and LH are produced by the same cell type, they are not always cosecreted, indicating that gonadotrophs must respond to unique signals that promote release of the appropriate hormone at the appropriate time. One determinant of gonadotroph response is the magnitude and pulsatility of GnRH signaling [10]. Under normal conditions, GnRH is released episodically, which elicits an appropriate response from the pituitary. If the interval between GnRH pulses is shortened, a phenomenon observed in many women with polycystic ovary syndrome (PCOS), the rate of LH release increases relative to that of FSH. If GnRH pulsatility is lost altogether and the pituitary is continuously exposed to high GnRH concentrations, the expression of gonadotrophic GnRH receptors is suppressed, the pituitary becomes less responsive to GnRH signaling, and FSH and LH release is inhibited. This observation has led to the development of GnRH agonists and analogs to treat conditions like precocious puberty and prostate cancer in which downregulation of gonadotrophic activity is desirable.

Adding to the complexity of the hypothalamus–pituitary–gonadal axis is the existence of different feedback inhibitory mechanisms in men and women. In men, testosterone and inhibin produced by the testes feedback on the pituitary and hypothalamus, inhibiting further FSH and LH release. In women, estradiol, progesterone, and inhibin produced by the ovaries provide negative feedback inhibition on the hypothalamus and pituitary, but this effect varies depending on the stage of the menstrual cycle. Estradiol inhibits FSH and LH release in the early follicular and luteal phases but actually provides positive feedback on the pituitary in the late follicular phase, promoting the LH surge required to trigger ovulation of the dominant ovarian follicle [11].

Following permanent cessation of ovarian activity at menopause, which typically occurs after approximately 45 years of age, plasma estradiol, progesterone, and inhibin fall to concentrations well below those observed in women of reproductive age. This is not a sudden change but rather occurs over a period of 4–5 years. This has important consequences for pituitary function, as the ovarian hormones that previously provided feedback inhibition have now been removed from the hypothalamus–pituitary–gonadal regulatory axis. This derepression of the pituitary leads to increased FSH production and substantially higher plasma FSH concentrations in postmenopausal women relative to premenopausal women at any stage of the menstrual cycle. Postmenopausal plasma LH concentrations are also increased relative to the follicular and luteal phases of the menstrual cycle but are generally lower than those observed in the ovulation phase of the menstrual cycle following the LH surge. Derepression of the pituitary following menopause may also result in the secretion of very small amounts of hCG. This modest hCG secretion from the pituitary often results in plasma concentrations above the upper reference limit for nonpregnant women of reproductive age and urine concentrations that are sufficiently high to generate positive qualitative test results. The prevalence of pituitary hCG secretion in postmenopausal women is unknown, but positive hCG test results obtained during a presurgical workup intended to exclude pregnancy before initiating the procedure may cause confusion among clinicians and their patients. In women who may be perimenopausal or postmenopausal, a pituitary source of hCG is most likely if the plasma concentration is <20 IU/L and plasma FSH is >45 IU/L. Suppression of plasma hCG following 2 weeks of estradiol supplementation further supports a pituitary source. If hCG remains elevated, hCG-producing malignancies, such as gestational trophoblastic disease, must be considered [12].

Measurement of FSH and LH in the evaluation of pituitary dysfunction is helpful in individuals with clinical signs and symptoms of hypogonadism. In children, hypogonadism presents as delayed puberty, while in adults, signs and symptoms include decreased libido, sexual dysfunction, menstrual abnormalities, and infertility. As discussed in greater detail in Chapter 43, Laboratory testing in reproductive disorders, of this book, low plasma concentrations of gonadal sex steroids can support a diagnosis of

hypogonadism. While men do not experience an abrupt cessation of testosterone production similar to that experienced by women at menopause, plasma testosterone concentrations decrease gradually with age. A plasma total testosterone concentration below the lower limit of the age-appropriate reference interval or a low free testosterone in the context of a normal total testosterone is consistent with male hypogonadism. In women, low plasma estradiol and progesterone concentrations are consistent with female hypogonadism, although these values must be interpreted in the context of the stage of the menstrual cycle when the specimen is collected. Relatively low estradiol and progesterone concentrations are expected in the early follicular phase, while, similarly, low concentrations in the mid-luteal phase suggest an underlying pathology.

Hypogonadism can be further subdivided into two groups with distinct mechanistic causes. Hypergonadotropic hypogonadism is characterized by gonadal failure (inability of the gonads to produce sex steroids in response to FSH and LH), while hypogonadotropic hypogonadism is defined by a central defect (inability of the hypothalamus to produce GnRH or of the pituitary to produce FSH and LH). Measurement of plasma or serum FSH and LH in the unstimulated state is often sufficient to distinguish between the two conditions, particularly if the concentrations of the two hormones are above the reference interval. Elevated FSH and LH in an unstimulated specimen establish the diagnosis of hypergonadotrophic hypogonadism and allow the clinician to eliminate other possible causes. On the other hand, because FSH and LH are secreted in a pulsatile fashion, a single unstimulated specimen with FSH and LH concentrations below the reference interval does not rule in hypogonadotropic hypogonadism. To avoid misinterpretation of a single low value, it is recommended to collect three serum or plasma samples with a 20−30-minute interval between each sample collection. FSH and LH can be measured either in each independent sample (three FSH and LH measurements) or in a mixed specimen created by pooling equal volumes of each of the three samples (one FSH and LH measurement). Low LH and FSH values observed under these conditions indicate insufficient pituitary production and strongly suggest a diagnosis of hypogonadotropic hypogonadism.

Stimulatory testing designed to provoke an FSH and LH response is generally unnecessary in adults but may help differentiate between hypogonadotropic hypogonadism and delayed puberty in children with low basal FSH and LH values. The GnRH stimulation test is conceptually similar to other provocative tests of pituitary function in that IV access is first obtained, followed by collection of a baseline sample for FSH and LH measurement. GnRH is then infused at an appropriate weight-based dose (2.5 μg/kg up to a maximum of 100 μg), and sequential specimens are collected up to 90 minutes after GnRH infusion at 15−20-minute intervals. Simpler variations of this procedure include subcutaneous injection of 100-μg GnRH followed by LH and FSH measurement at 40-minute postinjection or subcutaneous injection of the GnRH agonist leuprolide at 20 μg/kg up to 500 μg followed by LH and FSH measurement at 180-minute postinjection. Elevated LH and FSH after GnRH infusion support a diagnosis of delayed puberty, while low LH and FSH indicate hypogonadotropic hypogonadism. In contrast to the GH stimulation tests that define a "normal" response on the basis of a maximal GH concentration, GnRH stimulation tests are interpreted on the basis of a rise in FSH and LH relative to the patient's baseline values. It is also important to note that the GnRH stimulation test has limited diagnostic specificity, as there is substantial overlap in stimulated FSH and LH values between adolescents with hypogonadotropic hypogonadism and those with delayed puberty [13].

Hypergonadism is generally not caused by pituitary hyperfunction but is rather due to other pathophysiological conditions marked by excessive sex steroid production in peripheral tissues. These include androgen- or estrogen-secreting tumors derived from adrenal or gonadal tissue, PCOS, adult-onset congenital adrenal hyperplasia, or Cushing's syndrome. While these are serious health issues, they are beyond the scope of this chapter. Refer to Chapters 42, Laboratory testing in pregnancy, and Chapter 43, Laboratory testing in reproductive disorders, of this book for further discussion of FSH and LH in normal reproductive function and reproductive disorders.

Prolactin

Prolactin is the principal endocrine hormone that controls the initiation and maintenance of lactation. Prolactin is comprised of 199 amino acids and is produced and secreted by the anterior pituitary lactotroph (also referred to as mammotroph) cells. Other names for prolactin include lactogen, lactotropin, luteotropin, mammotropin, and lactogenic or luteogenic hormone. Dopamine is a neuroendocrine hormone produced in the hypothalamus and is the major regulator of prolactin (Fig. 39.11). Notably, dopamine suppresses prolactin production and release, which is dissimilar to other anterior pituitary hormones that exhibit stimulated release by hypothalamic regulatory hormones. Because of this inhibitory effect on prolactin release, dopamine may also be referred to as PRIH.

Approximately 85% of circulating prolactin is monomeric (23 kDa, also known as "little" prolactin) and is biologically active. Heterogenous forms of prolactin have also been detected in plasma, including dimeric prolactin (referred to as "big" prolactin) and polymeric forms (e.g., tetrameric, referred to as "big−big" prolactin). Select patients may also develop autoantibodies (e.g., IgG) that

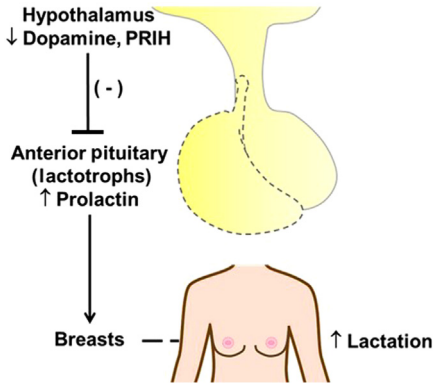


FIGURE 39.11 Prolactin regulatory axis. Prolactin production is under a constant state of repression by hypothalamic dopamine (prolactin release inhibiting hormone). Following removal of dopamine repression, prolactin is released into circulation where it stimulates differentiation of mammary alveoli and lactation.

bind to prolactin. These larger forms of prolactin are referred to as "macroprolactin." Macroprolactin has been suggested to have limited bioavailability, and can accumulate in circulation due to prolonged clearance [14,15]. The presence of macroprolactin does not reflect a pathologic process and is not associated with any specific disease state. However, it is important for laboratorians to recognize macroprolactin as a benign cause of elevated serum/plasma prolactin to spare the patient unnecessary evaluation and treatment. Biologically active monomeric prolactin binds to cell-surface receptors on target tissues (e.g., breast, ovaries, testes, prostate, kidney, and liver) and induces signal transduction for the modulation of gene transcription. With proper hormonal priming of breast tissue (e.g., estrogen), prolactin stimulates the differentiation of the mammary alveoli and subsequent synthesis of milk proteins, including casein and lactalbumin. Prolactin also plays an essential role in steroidogenesis and has been suggested to modulate immune system and metabolic functions.

Hyperprolactinemia (i.e., prolactin excess) is one of the most common hypothalamus–pituitary disorders encountered in clinical endocrinology practice. Signs, symptoms, and sequelae associated with hyperprolactinemia vary and may include amenorrhea or oligomenorrhea, galactorrhea, infertility, oligospermia, impotence, hypogonadism, and bone loss, or individuals may remain asymptomatic. Hyperprolactinemia can develop from several different physiological, pathophysiological, and pharmacological causes (Table 39.6). Physiologic stimuli known to stimulate prolactin release include coitus, pregnancy, lactation, sleep, stress, and exercise. During pregnancy, serum prolactin concentrations may increase 10-fold and reach concentrations of 150–300 ng/mL (150–300 μg/L) by term. Measurement of serum prolactin is not recommended in pregnant patients with prolactinomas (i.e., benign tumor) in order to minimize risk of potential misinterpretation and unnecessary testing triggered by prolactin concentrations greater than reference interval. Pathophysiologic prolactinomas, which are benign tumors (i.e., adenoma) that develop from lactotroph cells, cause nonpuerperal hyperprolactinemia and account for approximately 40% of all pituitary tumors [14]. Hypothalamic disease or interruption of the hypothalamus–pituitary portal system can occur due to several different etiologies and cause diminished dopamine suppression of prolactin secretion, resulting in hyperprolactinemia. In addition, prolactin can be released from the anterior pituitary in response to high concentration of TRH found with hypothyroidism. Several therapeutic drugs can stimulate

TABLE 39.6 Etiologies of hyperprolactinemia.

Physiological
• Coitus • Exercise • Lactation (nipple stimulation) • Pregnancy • Sleep • Stress
Pathological
Hypothalamus–pituitary stalk damage (dopamine/PRIH deficiency)
• Granulomas • Irradiation • Rathke's cyst • Trauma (pituitary stalk section and suprasellar surgery) • Tumors (craniopharyngioma and germinoma)
Pituitary
• Acromegaly • Prolactinoma • Secondary hypothyroidism (pathologically elevated TRH) • Surgery • Trauma
Other
• Chest wall disease • Chronic renal disease (impaired prolactin clearance) • Cirrhosis
Pharmacological
• Cholinergic agonist • Dopamine receptor blockers • Dopamine synthesis inhibitor • Estrogens (oral contraceptives and oral contraceptive withdrawal) • Serotonergic antagonists
Analytical
• Macroprolactinemia

TRH, Thyrotropin-releasing hormone.

prolactin release and create a biochemical picture of prolactinoma in an otherwise healthy individual. A careful patient history is needed to rule out the possibility that medications are not the cause for elevated prolactin.

The diagnosis of hyperprolactinemia is established based on laboratory testing for serum prolactin that show results greater than the reference interval [14]. Serum prolactin reference intervals are method-specific, though values are higher in females than in males and are generally <25 ng/mL (<25 μg/L). Determining the etiology of hyperprolactinemia is more challenging, however, and necessitates a comprehensive patient history, review of the patient's medications list, potential imaging studies (e.g., MRI of brain), and interpretation of the absolute concentration of measured prolactin in the clinical context. In patients with prolactinomas, serum prolactin concentrations generally parallel tumor size with prolactin elevations of >250 ng/mL. However, select drugs including risperidone and metoclopramide may cause prolactin elevations of >200 ng/mL in patients without clinical evidence of adenoma. Serum prolactin >500 ng/mL is diagnostic of a macroprolactinoma, which are macroadenomas >10 mm in diameter that secrete prolactin that may or may not include macroprolactin. Dynamic tests for prolactin secretion using TRH, L-dopa, nomifensine, and domperidone are not recommended.

Evaluation for potential high-dose hook interference

In patients with prolactinomas, serum prolactin concentrations generally parallel tumor size with prolactin elevations of >250 ng/mL. Prolactin-secreting macroadenomas can sometimes produce exceedingly high serum prolactin concentrations that can paradoxically cause measurement of falsely low prolactin concentrations by immunometric assays. In these situations, very high concentrations of prolactin saturate both the capture and detection antibodies in the assay, which subsequently blocks the formation of the capture antibody−prolactin−detection antibody "sandwich," resulting in falsely low prolactin results (i.e., high-dose hook interference). If clinical findings show the presence of a pituitary tumor and symptoms of prolactinoma but lower than expected serum prolactin concentration, an evaluation for potential falsely low serum prolactin measurement due to high-dose hook interference should be considered. Investigation of suspected high-dose hook interference involves serial dilutions (e.g., 10-, 100-, and 400-fold dilutions) of the serum sample in order to minimize the analytic artifact, with testing of these diluted samples using the same prolactin immunometric assay. If results show linear dilution and expected recoveries (e.g., ±20%) compared with the prolactin concentration in the neat sample (i.e., undiluted), then high-dose hook interference is ruled out. If significantly increasing concentrations of prolactin are found after dilution of the serum, then high-dose hook interference is observed and should be communicated with the clinical care team for appropriate patient management.

Evaluation for potential macroprolactinemia

In patients with asymptomatic hyperprolactinemia, assessment for potential presence of macroprolactin should be considered [16]. Immunoassay cross-reactivity with macroprolactin moieties is variable and method-dependent. Investigation of potential presence of macroprolactin can be performed by pretreating the serum sample with polyethylene glycol (PEG) followed by centrifugation that precipitates high molecular mass moieties. The supernatant containing monomeric prolactin is subsequently tested using the same prolactin immunoassay, and the result is compared with the prolactin concentration obtained in the neat sample (i.e., untreated, total prolactin). Of note, the dilution effect that occurs during sample pretreatment with PEG and the coprecipitation of monomeric prolactin must be taken into consideration. The precipitated prolactin concentration is calculated by subtracting the unprecipitated prolactin value (i.e., PEG-pretreated sample) from the total prolactin value. The percentage of precipitated prolactin to total prolactin is an indicator of the presence or absence of macroprolactin. When the fraction (i.e., percentage) of PEG-precipitated prolactin is ≤60% of total prolactin, the sample is considered negative for macroprolactin, and other causes for hyperprolactinemia should be evaluated. When the fraction of PEG-precipitated prolactin is >60% of total prolactin, the sample is considered positive for the presence of macroprolactin. While a 60% threshold for PEG-precipitated prolactin is described above for example purposes, this threshold is method-dependent and should be established using PEG-precipitated reference sera [17]. Careful interpretation of prolactin results in conjunction with clinical findings is needed, because the finding of macroprolactin does not exclude the possibility of concomitant presence of a prolactin-secreting pituitary adenoma and true hyperprolactinemia.

Posterior pituitary hormone pathophysiology and lab findings

Antidiuretic hormone

The major physiological functions of ADH [also referred to as arginine vasopressin or vasopressin] are to conserve free water and stimulate vasoconstriction, thus, ADH also plays key roles in maintaining blood pressure and osmolality of the extracellular space. ADH is a 9-amino acid peptide that

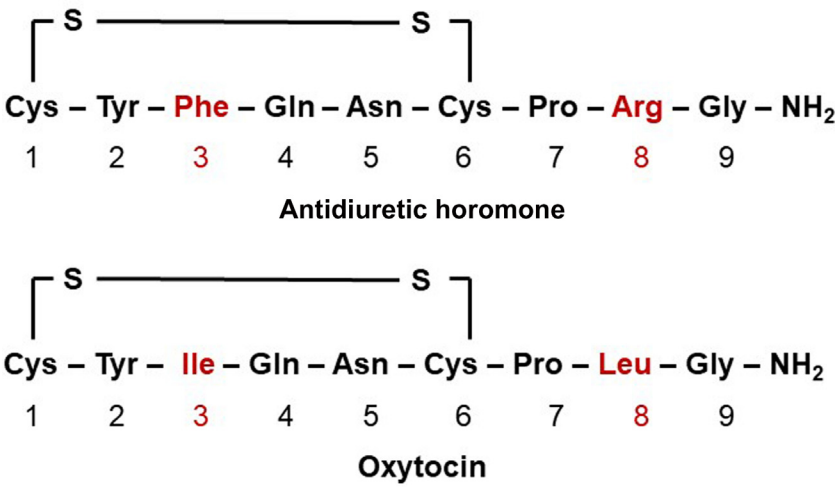


FIGURE 39.12 Amino acid sequences of antidiuretic hormone and oxytocin.

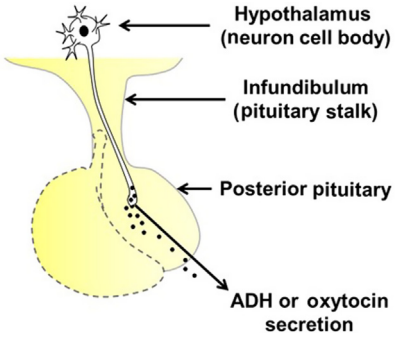


FIGURE 39.13 Posterior pituitary hormone secretion. Note that posterior pituitary hormones are synthesized in the hypothalamus and released by the posterior pituitary. *Adapted from W. Winter, N.S. Hardt, N.S. Harris, Disorders of the anterior and posterior pituitary.* Contemporary Practices in Clinical Chemistry*, third edn., 2016.*

is synthesized by neurosecretory cells in the hypothalamus (Fig. 39.12). Axons from neuron cell bodies originating in the hypothalamus transport ADH through the infundibulum to the posterior pituitary gland for storage (Fig. 39.13). A key regulator of ADH secretion into circulation by the posterior pituitary gland is the osmolality of the blood. Osmoreceptors located on hypothalamic neuron cell bodies respond to changes in plasma osmolality where an increase in plasma osmolality of $<1\%$ can stimulate ADH secretion. The secretion of ADH is also regulated by baroreceptors through a pressure−volume mechanism. High-pressure arterial baroreceptors of the aortic arch and carotid sinus, as well as low-pressure volume receptors in the pulmonary venous system and atria, respond to alterations in blood volume. Baroreceptors detection of decreased circulating blood volume or decreased blood pressure results in the stimulation of ADH secretion. Baroreceptor regulation of ADH secretion is relatively insensitive compared with osmoreceptors and requires a decrease of approximately 5% plasma volume, though upon activation elicits a potent response. While baroreceptors play a key role in maintaining plasma volume and pressure, this mechanism works in concert with the renin−angiotensin−aldosterone system, which is the major endocrine system regulating blood volume and pressure. Other stimuli for ADH secretion include pain, stressful conditions, and chemical agents (e.g., anesthetics, barbiturates, catecholamines, opiates, and prostaglandins). Altogether, daily water homeostasis is predominantly regulated by osmolality via osmoreceptor-mediated ADH secretion and its downstream effects, though this regulatory mechanism may be overridden by blood volume and pressure regulatory mechanisms in order to maintain normal plasma volume.

ADH binds to vasopressin receptors in order to mediate free water reabsorption and vasoconstriction. Vasopressin type 1 receptors are expressed on vascular smooth muscle cells and other tissues. The binding of ADH to type 1 vasopressin receptors on vascular smooth muscle cells promotes vasoconstriction. ADH also binds to vasopressin type 2 receptors expressed on collecting duct epithelial cells, which stimulates the kidney to reabsorb water in order to help maintain water homeostasis. The binding of ADH to vasopressin type 2 receptors signals the relocation of aquaporin-2 from an intracellular pool to the apical plasma membrane of collecting duct cells (Fig. 39.14). Aquaporin-2 is a water channel that, when localized in the luminal plasma membrane, increases the collecting ducts' permeability to water supporting free water reabsorption. Following water entry into collecting duct cells via aquaporin-2, water exits these cells into circulation via other water channels, including aquaporin-3 and aquaporin-4. Collectively, this physiological process promotes free water movement down an osmotic gradient. An incidental consequence of this renal reabsorption of water is concentrated urine and reduced urine volume. ADH can also bind to vasopressin type 3 receptors that are expressed on the anterior

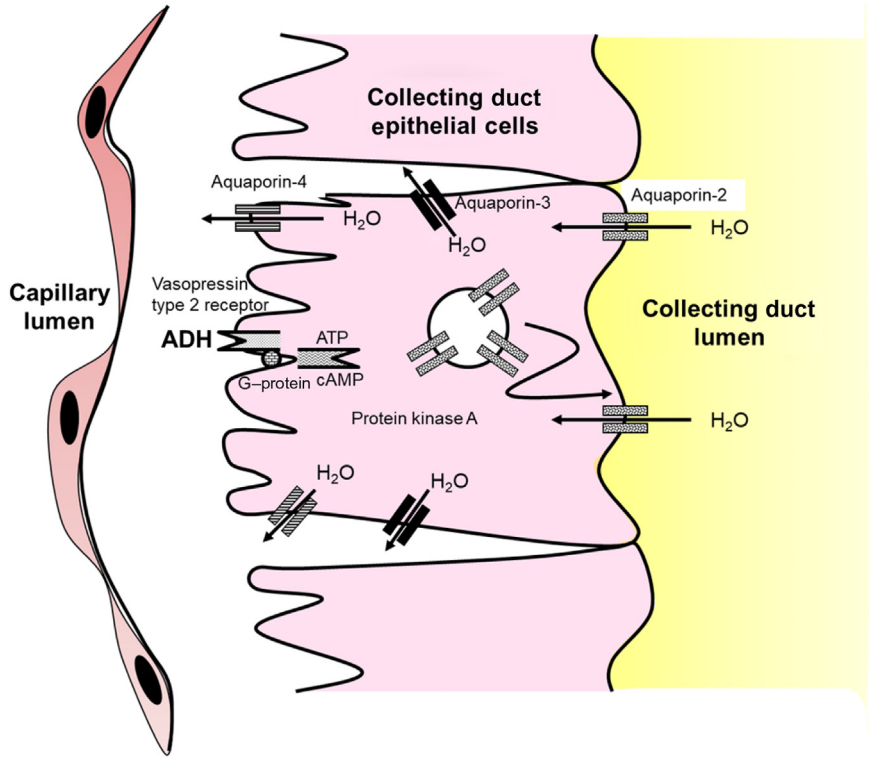


FIGURE 39.14 Mechanism of free water reabsorption stimulated by antidiuretic hormone. Antidiuretic hormone stimulates the relocation of aquaporin-2 into the apical plasma membrane of epithelial cells in the renal collecting duct, facilitating reabsorption of free water from the renal filtrate into circulating plasma. *Adapted from W. Winter, N.S. Hardt, N.S. Harris, Disorders of the anterior and posterior pituitary.* Contemporary Practices in Clinical Chemistry, *third edn., 2016.*

pituitary corticotrophs, resulting in the stimulation of ACTH secretion in response to high ADH concentration.

Syndrome of inappropriate antidiuretic hormone (SIADH) refers to autonomous production and secretion of ADH in the absence of known physiological stimuli. In this syndrome, plasma ADH concentration is "inappropriately" elevated relative to low plasma osmolality and to an increased or normal plasma volume. There are numerous potential etiologies for SIADH, including acute or chronic disorders of the central nervous system (CNS), ectopic ADH production by malignancy (e.g., small cell lung carcinoma), pulmonary disease, or a side effect of select drug therapies (ACE inhibitors, first generation sulfonylureas, second generation antipsychotics, and antidepressants). SIADH in patients with unrestricted fluid intake causes inappropriate antidiuresis, because the high ADH concentrations stimulate excessive reabsorption of free water by the kidneys, leading to water retention with reduced urine output. This increase in intravascular volume can cause hemodilution accompanied by dilutional hyponatremia and a low plasma osmolality, features that should normally inhibit ADH secretion. Volume expansion also suppresses the renin–angiotensin–aldosterone system, resulting in urinary sodium loss with elevated urine osmolality and urinary sodium concentration. Despite modest volume expansion in patients with SIADH, edema is not commonly observed. Clinical manifestations of hyponatremia may arise and vary from mild to severe (e.g., weakness, apathy, lethargy, coma, and seizure); however, there are no clinical signs or symptoms that are specific for SIADH.

Diabetes insipidus (DI) refers to a pathological condition due to inadequate production or action of ADH. Signs, symptoms, and sequelae associated with DI include polyuria with excretion of large volumes (>10 L/day) of dilute urine due to excess urinary free water loss that consequently causes potential hypernatremia and associated elevated plasma osmolality. In patients with intact thirst sensors (refer to Chapter 37: Water and electrolyte balance, of this book for more information on this topic), the elevated plasma osmolality stimulates thirst, and patients may report polydipsia in addition to polyuria. In such cases, excess intake of free water can help compensate for the excessive urinary loss of water, and patients may exhibit only mild hypernatremia. If free water is unavailable or the patient has dysfunction of a normal thirst mechanism, dehydration can occur due to excessive urinary loss of water, leading to a decline in blood pressure and tachycardia.

DI is categorized as central DI or nephrogenic DI, depending on the mechanism of dysfunction. Central DI can arise from any destructive hypothalamic lesion, infundibular lesion, or posterior pituitary lesion that impedes the production or secretion of ADH, resulting in ADH deficiency (Fig. 39.15). There are numerous potential etiologies of central DI, including CNS abnormalities such as

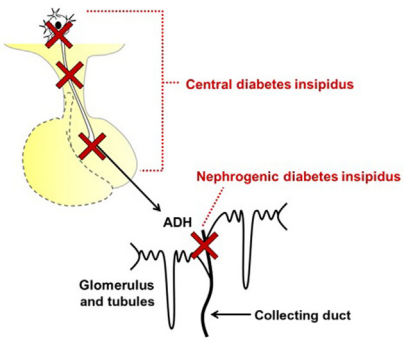


FIGURE 39.15 Pathophysiological mechanisms of central and nephrogenic diabetes insipidus. Central diabetes insipidus is caused by a defect in antidiuretic hormone production and/or secretion, whereas nephrogenic diabetes insipidus is caused by partial or complete resistance to antidiuretic hormone by the kidneys. *Adapted from W. Winter, N.S. Hardt, N.S. Harris, Disorders of the anterior and posterior pituitary.* Contemporary Practices in Clinical Chemistry*, third edn., 2016.*

hypothalamic or pituitary tumors, stroke, head trauma, infiltrative disease (e.g., sarcoidosis), and others. Contrastingly, DI arising from partial or complete resistance to ADH action due to renal tubular dysfunction or disease is referred to as "nephrogenic DI." Nephrogenic DI is a rare kidney disorder that can be inherited or acquired. One form of hereditary nephrogenic DI is an X-linked recessive disorder that causes an inactivating mutation of the vasopressin type 2 receptor (AVPR2 gene), leading to ADH resistance. Other congenital forms of nephrogenic DI include autosomal recessive or autosomal dominance loss-of-function mutations of aquaporin-2 (AQP2 gene), leading to dysfunctional water channels and reduced water reabsorption despite ADH stimulation. Acquired forms of nephrogenic DI are more commonly observed than hereditary nephrogenic DI. A common cause of acquired nephrogenic DI is patient exposure to drugs, notably chronic use of lithium, as well as various antimicrobial drugs (e.g., amphotericin B, rifampin, and methicillin) and antineoplastic drugs (e.g., cisplatin and ifosfamide). Moreover, nephrogenic DI can manifest due to any form of renal tubular injury accompanied by impaired water reabsorption (e.g., polycystic kidney disease, medullary cystic kidney disease, chronic pyelonephritis, obstructive uropathy, sickle cell nephropathy, and renal amyloidosis).

Routine laboratory testing together with a comprehensive patient history and physical examination are needed to diagnose SIADH and DI. Testing for ADH itself is not usually needed for diagnostic purposes, because ADH excess or deficiency is evident by clinical assessment of fluid balance (e.g., blood pressure and heart rate) and changes in serum and urine osmolalities, serum and urine sodium concentrations, and urine volume. Measurement of urine specific gravity is a simple screening test that helps assess the kidney's ability to concentrate urine and can provide insight into a patient's urine osmolality. Urine specific gravity is the ratio of the mass of a urine sample compared with the mass of an equal volume of water. Urine specific gravity is typically between 1.002 and 1.030 under normal physiological conditions. In general, patients with SIADH have a high urine specific gravity, whereas patients with DI typically have a low urine specific gravity value. The urine specific gravity test should not be performed in patients exhibiting proteinuria, hematuria, glycosuria, or excretion of other osmotically active compounds, such as contrast dyes. In these cases, a urine osmolality test is a better measure of a patient's ability to concentrate urine.

If a patient has a history of polyuria and polydipsia with an abnormal urine specific gravity result, additional testing should be performed, including serum and urine osmolalities, as well as a basic metabolic panel. Osmolality can be measured directly using freezing point depression. Additionally, serum osmolality (Osm_s) can be estimated using measured concentrations of the predominant osmotic substances in plasma including sodium (Na^+), blood urea nitrogen, and glucose. The following formulas are commonly used for estimating Osm_s:

$$\begin{aligned}\text{Serum Osm}_s = &\left(1.86 \times \left[\text{Na}^+\right](\text{mmol/L})\right) \\ &+ \left([\text{BUN}](\text{mg/dL}) \div 2.8\right) \\ &+ \left([\text{glucose}](\text{mg/dL}) \div 18\right) + 9\end{aligned} \tag{39.1}$$

or

$$\begin{aligned}\text{Serum Osm}_s = &\left(2 \times \left[\text{Na}^+\right](\text{mmol/L})\right) \\ &+ \left([\text{BUN}](\text{mg/dL}) \div 2.8\right) \\ &+ \left([\text{glucose}](\text{mg/dL}) \div 18\right)\end{aligned} \tag{39.2}$$

In Eq. (39.1), the constant 1.86 is the osmotic coefficient of sodium (0.93) multiplied by 2 to account for the dissociated anion paired with sodium. The osmotic coefficient is <1.0 due to incomplete electrolyte dissociation at typical plasma concentrations. The constant 9 accounts for the other osmotically active plasma components, including potassium, calcium, and other electrolytes, and proteins found at lower molar concentrations. Eq. (39.2) simplifies the calculation but often results in a higher calculated osmolality relative to Eq. (39.1). Under typical physiological conditions, the difference between measured and calculated Osm_s (osmolal gap) is expected to be <10 mOsm/kg, and a larger gap suggests the presence of unmeasured osmotically active substances.

The calculation of urine osmolality is conceptually similar but includes potassium and associated anions:

$$\begin{aligned}\text{Urine Osm} = &\left(2 \times \left[\text{Na}^+ + K^+\right](\text{mmol/L})\right) \\ &+ \left([\text{urea nitrogen}](\text{mg/dL}) \div 2.8\right) \\ &+ \left([\text{glucose}](\text{mg/dL}) \div 18\right)\end{aligned} \tag{39.3}$$

The urine osmolal gap typically falls between 10 and 100 mOsm/kg and can be used as a surrogate marker of urine ammonium (NH_4) excretion, but this is rarely done in routine clinical practice.

Serum and urine osmolality results vary depending on the patient hydration status, though, under normal physiological conditions, the reference intervals are typically between 275−295 and 300−900 mOsm/kg, respectively. The ratio of urine osmolality to Osm_s is typically 1−3 under normal physiological conditions, reflecting a wide range of urine osmolalities.

In SIADH with inappropriate antidiuresis and excessive free water retention, there are aberrations in several routine clinical laboratory values. Patients with SIADH exhibit reduced Osm_s (e.g., <275 mOsm/kg) and hyponatremia (e.g., <135 mEq/L). In the setting of hyponatremia and low Osm_s, the normal physiological response is to maximally dilute urine, leading to a urine osmolality of 50−80 mOsm/kg. However, in patients with SIADH, the kidneys inappropriately concentrate urine (e.g., urine osmolality >100 mOsm/kg) and have reduced urine output. The urine osmolality may not exceed the Osm_s in SIADH but will show inappropriate urine concentration relative to Osm_s. This pattern of lab values together with clinical evidence of volume expansion supports the diagnosis of SIADH. The water loading test is a dynamic test, which is sometimes used to help differentiate potential causes of hyponatremia; however, this test is contraindicated in patients with SIADH, because excess fluid intake can exacerbate hyponatremia with potential serious adverse outcomes.

The diagnosis of DI involves routine clinical laboratory testing to evaluate potential excessive urinary free water loss. A low urine osmolality of <300 mOsm/kg and elevated Osm_s of >300 mOsm/kg (or hypernatremia is present) support the diagnosis of DI. A water deprivation test with desmopressin challenge may also be performed to differentiate whether the etiology of DI is central or nephrogenic, or be used when the diagnosis of DI is uncertain.

Water deprivation test

A water deprivation test typically begins in the morning at which point the patient is given water to drink (e.g., 20 mL/kg) over a 15−30-minute period, and food and water are withheld throughout the duration of the test. Serial collections of blood and urine samples (e.g., baseline, hourly, or bi-hourly) are performed with measurement of serum and urine osmolalities, respectively. A rising trend in Osm_s while urine osmolality remains low during the water deprivation test is an indicator of DI (Fig. 39.16). The water deprivation testing continues for 8−10 hours unless the lab values establish the diagnosis of DI prior to completion, or test termination is prompted by mental status changes or hypotension. In order to further determine if DI is from a central or nephrogenic etiology, a desmopressin challenge test can be performed during the clinical water deprivation procedure.

Desmopressin challenge test

Desmopressin (i.e., desamino-D-arginine, DDAVP) is a biologically active analog of endogenous ADH (vasopressin) that can be administered during the water deprivation test followed by measurement of serum and urine osmolalities. In patients with central DI, the kidneys respond to desmopressin stimulation by increasing water reabsorption and subsequently have more concentrated urine with an increase (e.g., >doubling) in urine osmolality (e.g., Fig. 39.16A). Conversely, patients with nephrogenic DI exhibit minimal renal response to exogenous desmopressin, and urine osmolality remains low (Fig. 39.16B).

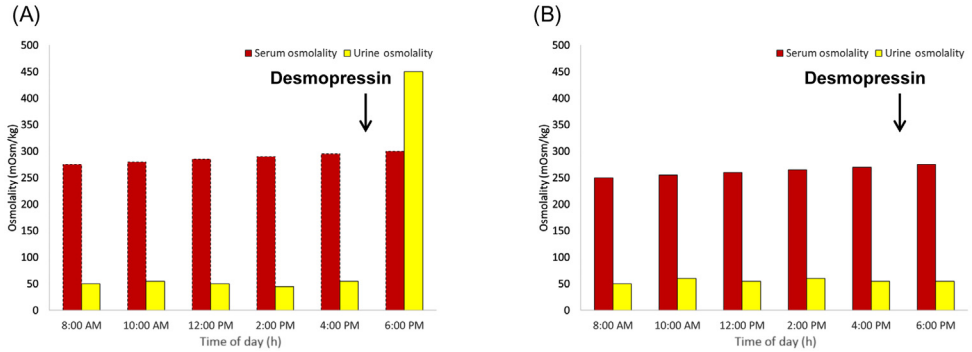


FIGURE 39.16 Expected serum and urine osmolality responses to water deprivation testing with desmopressin challenge in the setting of diabetes insipidus. Inappropriately, low urine osmolality and elevated plasma osmolality following water restriction are consistent with diabetes insipidus. The desmopressin challenge is an additional provocative test to help differentiate central versus nephrogenic diabetes insipidus. (A) Marked renal response with an increase in urine osmolality following desmopressin administration is consistent with central diabetes insipidus. (B) Minimal change in urine osmolality following desmopressin administration is consistent with nephrogenic diabetes insipidus.

Oxytocin

Oxytocin plays major roles in parturition and lactation. Similar to ADH, oxytocin is a nonapeptide (9-amino acids) that is synthesized in the hypothalamus and packaged into secretory granules that migrate down nerve exons into the posterior pituitary lobe for storage (Figs. 39.12 and 39.13). The release of oxytocin is controlled primarily by neuroendocrine reflex, leading to hypothalamus−posterior pituitary stimulation. The neuroendocrine reflex is activated by neural stimuli originating in the uterus and genital tract (distension of the vagina and cervix), as well as the nipples (i.e., suckling). Upon neural stimulation, oxytocin is released from the posterior pituitary via calcium-dependent exocytosis and circulates in plasma in an unbound form. Oxytocin binds to G-protein coupled cell-surface receptors on smooth muscle cells, leading to an increase in intracellular Ca^{2+} concentration, which promotes contraction of the smooth muscle cells of the uterine myometrium during parturition, as well as the contractile elements of the mammary alveoli associated with milk letdown during nursing. Oxytocin stimulation of smooth muscle contraction is dependent on estrogen priming of the uterus and mammary alveoli. Exogenous oxytocin (also referred to as Pitocin) can be administered for patient care in select settings. More specifically, exogenous oxytocin is FDA-approved for therapeutic use for the induction of labor in patients who have medical indications for induction (e.g., preeclampsia and premature rupture of membranes), stimulation, or reinforcement of uterine contraction, and as an adjunctive therapy in the management of incomplete or inevitable abortion. Laboratory testing for the measurement of plasma oxytocin concentration is not routinely performed, in part, because of limited evidence of physiological relevance to human reproductive disorders and limited clinical usefulness in clinical decision-making for patient care.

References

[1] B.R. Javorsky, et al., Hypothalamus and the pituitary gland, in: D. Gardner, D. Shoback (Eds.), McGraw-Hill Publishers, New York, *Greenspan's Basic and Clinical Endocrinology*, 2018, pp. 69−118.

[2] A. Tjörnstrand, et al., The incidence rate of pituitary adenomas in western Sweden for the period 2001−2011, Eur. J. Endocrinol. 171 (4) (2014) 519−526.

[3] V. Gounden, Y.D. Rampursat, I. Jialal, Secretory tumors of the pituitary gland: a clinical biochemistry perspective, Clin. Chem. Lab. Med. (2018).

[4] D.D. Hamilton, B.A. Cotton, Cosyntropin as a diagnostic agent in the screening of patients for adrenocortical insufficiency, Clin. Pharmacol. 2 (2010) 77−82.

[5] G. Patti, et al., Central adrenal insufficiency in children and adolescents, Best. Pract. Res. Clin. Endocrinol. Metab. 32 (4) (2018) 425−444.

[6] P. Cohen, et al., Consensus statement on the diagnosis and treatment of children with idiopathic short stature: a summary of the Growth Hormone Research Society, the Lawson Wilkins Pediatric Endocrine Society, and the European Society for Paediatric Endocrinology Workshop, J. Clin. Endocrinol. Metab. 93 (11) (2008) 4210−4217.

[7] S.M. Shalet, et al., The diagnosis of growth hormone deficiency in children and adults, Endocr. Rev. 19 (2) (1998) 203−223.

[8] L.A. Wetterau, The pros and cons of sex steroid priming in growth hormone stimulation testing, J. Pediatr. Endocrinol. Metab. 25 (11−12) (2012) 1049−1055.

[9] I.M. Holdaway, C. Rajasoorya, Epidemiology of acromegaly, Pituitary 2 (1) (1999) 29−41.

[10] R. Tsutsumi, N.J. Webster, GnRH pulsatility, the pituitary response and reproductive dysfunction, Endocr. J. 56 (6) (2009) 729−737.

[11] S.M. Moenter, Z. Chu, C.A. Christian, Neurobiological mechanisms underlying oestradiol negative and positive feedback regulation of gonadotrophin-releasing hormone neurones, J. Neuroendocrinol. 21 (4) (2009) 327−333.

[12] A.M. Gronowski, et al., Use of serum FSH to identify perimenopausal women with pituitary hCG, Clin. Chem. 54 (4) (2008) 652−656.

[13] J. Harrington, M.R. Palmert, Clinical review: distinguishing constitutional delay of growth and puberty from isolated hypogonadotropic hypogonadism: critical appraisal of available diagnostic tests, J. Clin. Endocrinol. Metab. 97 (9) (2012) 3056−3067.

[14] S. Melmed, et al., Diagnosis and treatment of hyperprolactinemia: an Endocrine Society clinical practice guideline, J. Clin. Endocrinol. Metab. 96 (2) (2011) 273−288.

[15] V. Richa, G. Rahul, A. Sarika, Macroprolactin; a frequent cause of misdiagnosed hyperprolactinemia in clinical practice, J. Reprod. Infertil. 11 (3) (2010) 161−167.

[16] M. Fahie-Wilson, In hyperprolactinemia, testing for macroprolactin is essential, Clin. Chem. 49 (9) (2003) 1434−1436.

[17] C.R. McCudden, J.L. Sharpless, D.G. Grenache, Comparison of multiple methods for identification of hyperprolactinemia in the presence of macroprolactin, Clin. Chim. Acta 411 (3−4) (2010) 155−160.

Chapter 40

Laboratory evaluation of thyroid function

David C. Lin[1] and Joely A. Straseski[2]

[1]Department of Pathology, Ann & Robert H. Lurie Children's Hospital of Chicago, Northwestern University Feinberg School of Medicine, Chicago, IL, United States, [2]Department of Pathology, University of Utah School of Medicine, ARUP Laboratories Salt Lake City, UT, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Outline the steps involved in thyroid hormone synthesis, including the roles of the following: thyroid-stimulating hormone (TSH), thyroglobulin, pendrin, Na^+/I^- symporter, and thyroperoxidase.
- Explain the roles of thyrotropin-releasing hormone, TSH, thyroxine, and triiodothyronine in the regulation of the hypothalamus–pituitary–thyroid axis.
- Compare and contrast the relative concentrations and binding characteristics of the predominant binding proteins of thyroid hormones.
- List the differences between Graves' disease and Hashimoto's thyroiditis in terms of: (1) clinical features; (2) biochemical changes; and (3) presence of specific autoantibodies.
- Define the difference between direct and indirect measurements of free thyroid hormones, as well as list the clinical scenarios in which direct testing may be preferred.

Thyroid: anatomy, physiology, and biochemistry

The thyroid gland, or simply the thyroid, is located in the lower anterior of the neck, just below the larynx and in front of the trachea. It is an "H"-shape organ that consist of two lobules, bridged by a band of thyroid tissue called the isthmus. The thyroid is one of the largest endocrine glands in the body, and is responsible for the production and secretion of thyroid hormones, which are vital for normal growth, development, and metabolism. Thyroid hormones act on virtually all organ systems and can trigger an array of biological effects (summarized in Table 40.1).

Thyroid hormone regulation

Hypothalamus–pituitary–thyroid axis

Secretion of thyroid hormones is regulated by the hypothalamus–pituitary–thyroid axis (HPTA) (Fig. 40.1). This axis involves several key regulatory hormones, including thyrotropin-releasing hormone (TRH) and thyroid-stimulating hormone (TSH), also referred to as thyrotropin. TRH is a 3-amino acid-long peptide chain (glutamic acid-histidine-proline), whereas TSH is a glycoprotein that consists of two subunits: an alpha subunit of 89 amino acids and a beta subunit of 112 amino acids. The alpha subunit of TSH is identical to that found in luteinizing hormone (LH), follicle-stimulating hormone (FSH), and human chorionic gonadotropin (hCG).

In the HPTA, the hypothalamus produces TRH, which enters the circulation and reaches the anterior pituitary via a network of blood vessel connections known as the hypophyseal portal system. TRH binds to a class of G-protein-coupled receptors (thyrotropin-releasing hormone receptor) found on endocrine cells, called thyrotrophs, in the anterior pituitary. This triggers the thyrotrophs to secrete TSH. TSH then binds to another G-protein-coupled receptor found on the membrane of the thyroid cells, activates a signal cascade involving adenylate cyclase, and causes an increase in intracellular cyclic adenosine monophosphate (cAMP) concentration.

Post TSH-binding, the increase of cAMP concentration in the thyroid cells results in: (1) thyroid cell proliferation and hypertrophy; and (2) synthesis and secretion of the thyroid hormones 3, 5, 3', 5' -tetraiodothyronine [thyroxine (T_4)] and 3, 5, 3'-triiodothyronine [triiodothyronine (T_3)] by the follicular cells. T_4 and T_3 secreted by the thyroid, as well as T_3 formed from T_4 in the periphery (see the next section), all provide negative feedback to the hypothalamus and pituitary to inhibit further secretion of TRH and TSH, respectively (discussed further at the end of this section).

Thyroid hormone synthesis

The thyroid gland is comprised of two types of cells—follicular and parafollicular ("C") cells. The follicular and C

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00040-5

TABLE 40.1 Systemic effects of thyroid hormones.

Systems affected	Effects
Nervous system	• Maturation and development of the CNS • ↑ sensitivity to catecholamines (via induced expression of catecholamine receptors)
Musculoskeletal system	• ↑ linear growth (via promotion of hypertrophic differentiation of the growth plate chondrocytes)

Metabolism	• ↑ basal metabolic rate[a]
	  ◦ ↑ oxygen consumption[b]
	  ◦ ↑ body heat production/thermogenesis[c]
	• ↑ carbohydrate metabolism
	  ◦ ↑ gluconeogenesis
	  ◦ ↑ glycogenolysis
	  ◦ ↑ glucose absorption
	• ↑ lipid metabolism
	  ◦ ↑ lipolysis
	• ↑ protein metabolism (overall more catabolic state)

Cardiovascular system	• ↑ cardiac output • ↑ blood flow • ↑ heart rate • ↑ sensitivity to catecholamines (leading to ↑ heart rate)

↑, Increased; *CNS*, central nervous system.

[a] *Metabolic rate at rest.*
[b] *In most tissues, via increased synthesis and activity of Na^+/K^+ adenosine triphosphatase.*
[c] *In part due to increased fuel oxidation and production of adenosine triphosphate.*

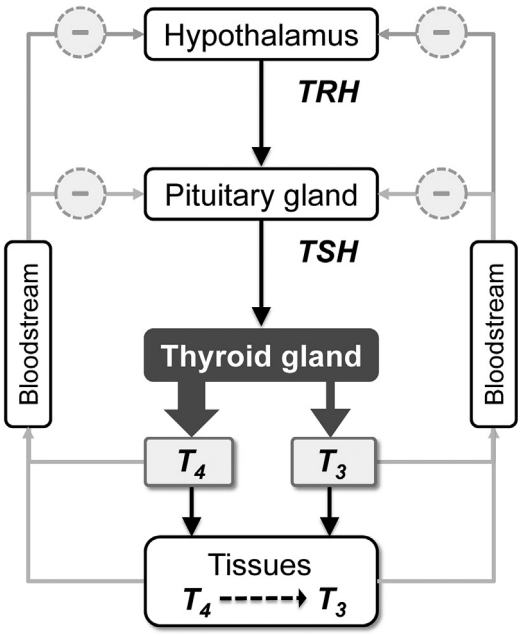


FIGURE 40.1 Hypothalamus–pituitary–thyroid axis. Thyrotropin-releasing hormone from the hypothalamus stimulates thyroid-stimulating hormone release from the pituitary gland. Thyroid-stimulating hormone then initiates thyroid hormone release from the thyroid gland, with more thyroxine produced than triiodothyronine. Triiodothyronine is predominantly produced via conversion of thyroxine within peripheral tissues. Triiodothyronine and thyroxine both provide negative feedback to the pituitary (predominantly) and hypothalamus (lesser) to halt further production. T_3, Triiodothyronine; T_4, thyroxine; *TRH*, thyrotropin-releasing hormone; *TSH*, thyroid-stimulating hormone; −, negative feedback.

cells are capable of producing thyroid hormones and calcitonin, respectively. The secretory units of the thyroid gland are called follicles (Fig. 40.2A). Each follicle consists of follicular cells surrounding the colloid, a gelatinous, amorphous material comprising newly made thyroid hormone bound to thyroglobulin (Tg). The follicles are separated by the interstitium, which is interspersed with capillaries, venules, and C cells.

Generation of thyroid hormones requires two main "ingredients": tyrosine (from Tg) and iodine (from the diet). Tg acts as the "platform" upon which these two ingredients combine to form the final thyroid hormone products. Given the central role of Tg, its production is compulsory to the remaining steps in thyroid hormone synthesis. Additionally, iodine must be transported to the appropriate thyroid region and converted into the correct chemical state before thyroid hormone "manufacturing" can take place. Thyroid hormones are synthesized through a complicated and unique series of steps occurring within the thyroid follicles, which are regulated by TSH. These steps are outlined below, with the numbers of each step corresponding to Fig. 40.2B.

1) Thyroglobulin synthesis

Tg, previously known as colloid antigen, is a 660-kDa glycoprotein that is only found within the thyroid. It is produced by the follicular cells and stored within the colloid. Tg is synthesized from tyrosine molecules by the

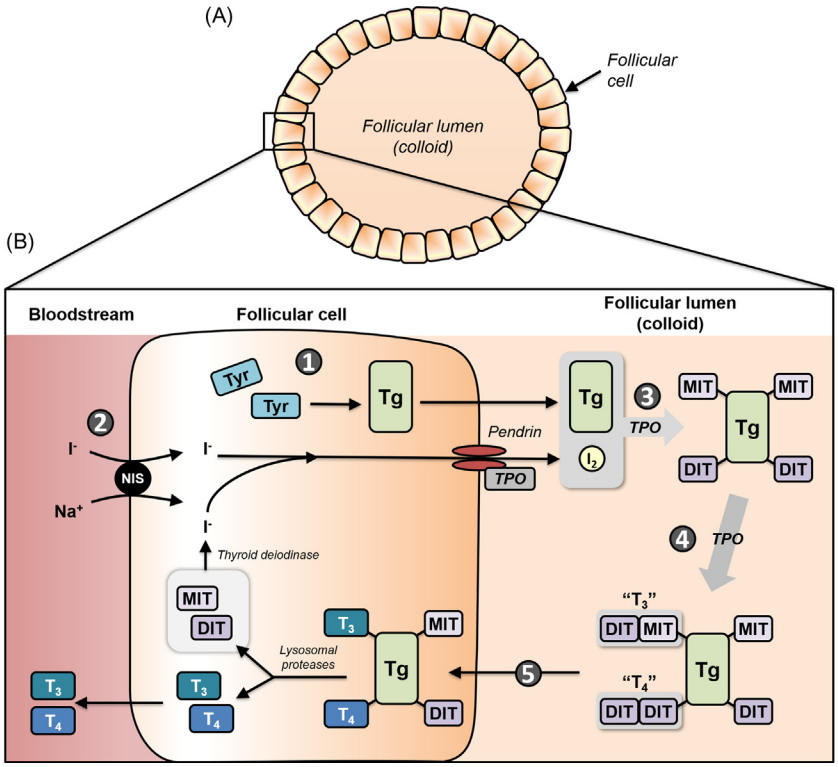


FIGURE 40.2 Anatomy and biochemistry involved in thyroid hormone synthesis. (A) General structure of the thyroid follicle, with follicular cells encompassing a lumen filled with colloid. (B) Major steps involved in thyroid hormone synthesis. Numbers correspond to steps detailed within the text (see "Thyroid hormone synthesis" section). *DIT*, Diiodotyrosine; I^-, iodide; I_2, iodine; *MIT*, monoiodotyrosine; *NIS*, Na^+/I^- symporter; T_3, triiodothyronine; T_4, thyroxine; *Tg*, thyroglobulin; *TPO*, thyroid peroxidase; *Tyr*, tyrosine.

follicular cells, packaged in secretary vesicles, and released into the follicular lumen via exocytosis. It is responsible for the majority of total protein and total protein synthesis within the thyroid gland. Although Tg is tyrosine-rich, only a small portion of the tyrosine residues can be iodinated (approximately 20%) and are involved in thyroid hormonogenesis.

2) Iodine transportation and processing

Iodine (I_2) is an essential component in thyroid hormone synthesis. Dietary sources, such as seafood and iodized table salt, contain the anionic form of the iodine atom, iodide (I^-), which needs to be absorbed and converted into I_2 before it can be utilized to generate thyroid hormones. I^- is absorbed through the gastrointestinal tract and distributed in the extracellular fluid (ECF). From the ECF, I^- can be excreted in the urine, or transported into the thyroid follicular epithelial cells via the sodium–iodide cotransporter (Na^+/I^- symporter). These cotransporters are found at the basal membrane of follicular cells and facilitate transport of I^- along with a charge-balancing Na^+ ion. I^- is then transported into the follicular lumen via an anion transporter protein called pendrin, which is present at the apical membrane of the follicular cells. An enzyme called thyroid peroxidase (TPO), or thyroperoxidase, then further oxidizes I^- to I_2.

3) Monoiodo tyrosine and diiodotyrosine formation

Inside the follicular lumen, I_2 binds to tyrosine residues of Tg to form monoiodo tyrosine (MIT) or diiodotyrosine (DIT) (see Fig. 40.3A). MIT has one iodine coupled to it, whereas DIT has two. The coupling reaction between Tg and I_2, sometimes referred to as "organification," is catalyzed by TPO. In the presence of excess I^-, organification is inhibited in order to protect against excess thyroid hormone synthesis. This is known as the Wolff–Chaikoff effect, and has historically been used as the basis for hyperthyroidism treatments.

4) Triiodothyronine and thyroxine formation

T_3 and T_4 are formed from MITs and DITs within the colloid. T_3 is formed when one MIT is coupled to one DIT, whereas T_4 is formed when two DITs are coupled (see Fig. 40.3B). Reverse triiodothyronine (rT_3) is formed similarly to T_3 but in opposite spatial configuration. The aforementioned coupling reactions are also mediated by TPO. The iodinated Tg (i.e., Tg with T_3, T_4, MIT, and/or DIT attached) serves as the I_2 storage compartment in the thyroid and accounts for approximately 90% of total body I_2 concentration (6000–12,000 μg) [1]. This is an important physiological feature, because thyroid hormones can affect numerous biological functions (Table 40.1).

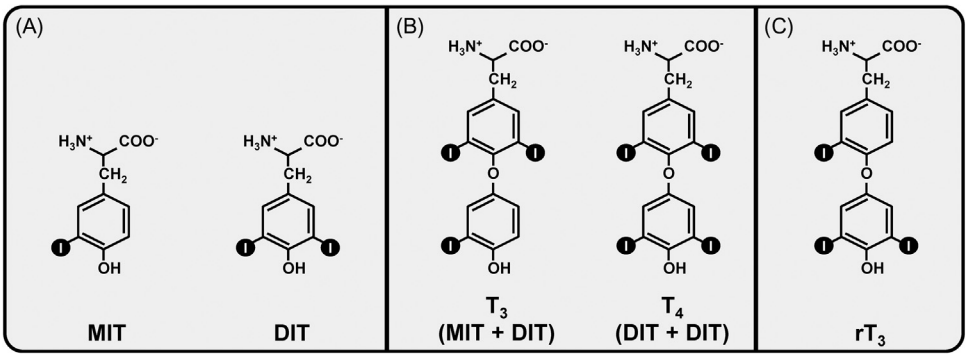


FIGURE 40.3 Chemical structure of thyroid hormones and their precursors. The phenol ring positioned nearest the amino acid terminal is referred to as the alpha ring, and the ring located on the hydroxyl terminal is the beta ring. *DIT*, Diiodotyrosine; *MIT*, monoiodotyrosine; *rT3*, reverse triiodothyronine; *T3*, triiodothyronine; *T4*, thyroxine; I, iodine atoms.

As such, having a pool of iodinated Tg allows the body to readily secrete thyroid hormone when needed. It also explains the transient hyperthyroidism that may be observed in the initial stages of hypothyroid disease; acute autoimmune destruction of the thyroid gland results in rapid release of these stored hormones (referred to as "Hashitoxicosis").

5) Thyroid hormone release and monoiodo tyrosine/diiodotyrosine recycling

Release of thyroid hormone is tightly controlled by TSH and the HPTA. Furthermore, the thyroid gland is capable of recycling the individual components of thyroid hormone synthesis, such as MIT and DIT. These features, along with the readily available iodinated Tg pool, allow for an effective thyroid hormone "factory," where: (1) production is tightly regulated; (2) unconsumed components are recycled to reduce waste; and (3) stored supply is available on demand.

When the thyroid is stimulated via TSH, Tg is taken up from the colloid into follicular cells via pinocytosis. Follicular lysosomes fuse with the endosome and lysosomal proteases digest Tg to release T_3, T_4, MIT, and/or DIT. T_3 and T_4 are transported out of the follicular cells and into the circulation. Due to their lipophilic nature, T_3 and T_4 must be transported in the circulation via binding proteins. MIT and DIT are deiodinated by thyroid deiodinase and release their I^-, which is then recycled for future production of thyroid hormones.

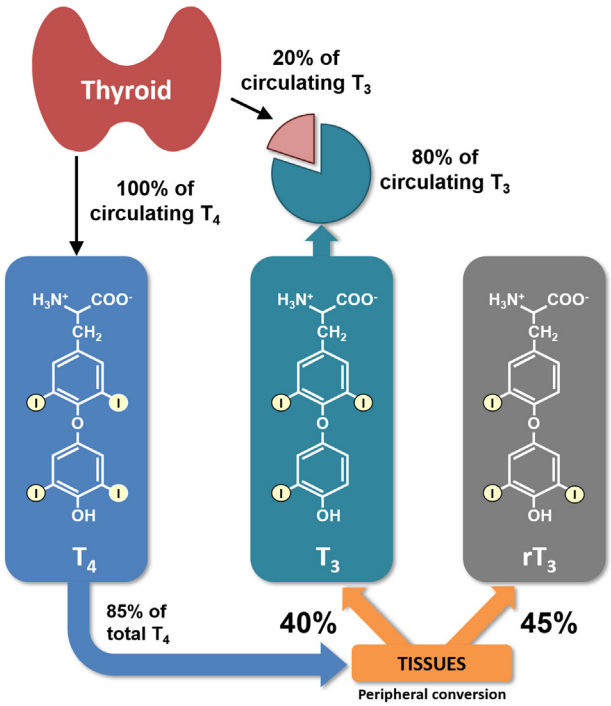


FIGURE 40.4 Relative secretion of thyroxine and triiodothyronine by the thyroid gland and further conversion of thyroxine into triiodothyronine or reverse triiodothyronine. The thyroid gland is the only source of circulating thyroxine, whereas the majority of triiodothyronine is produced via conversion of thyroxine at the peripheral tissues. Approximately equal amounts of triiodothyronine and reverse triiodothyronine are produced during thyroxine conversion. Regulation of these percentages can effectively regulate systemic thyroid functionality, as triiodothyronine is the biologically active hormone form and reverse triiodothyronine has no biological activity. *T3*, Triiodothyronine; *T4*, thyroxine; *rT3*, reverse triiodothyronine.

Thyroxine versus triiodothyronine

In general, the thyroid gland secretes more T_4 than T_3. The majority of T_3 is instead generated within the peripheral tissue by deiodination of T_4 (Fig. 40.4). Despite being produced at lower quantity than T_4 by the thyroid, T_3 is considered the more biologically active hormone, whereas T_4 is considered a prohormone. In the circulation, T_4 is relatively more protein-bound than T_3 (99.97% of T_4 versus 99.7% of T_3 is protein-bound). This, in turn, contributes to T_4's longer half-life and, therefore, higher concentration (Table 40.2). While the total T_4 concentration is approximately 50 times that of total T_3, the

TABLE 40.2 Concentrations of total and free thyroid hormones in the circulation.

Thyroid hormone	Total concentration	Free hormone percent	Free hormone concentration	Typical half-life
Thyroxine	4–12 μg/dL	~0.03%	0.8–2.7 ng/dL	6–7 days
Triiodothyronine	70–170 ng/dL	~0.3%	0.2–0.48 ng/dL	1–1.5 days

concentration of circulating free, unbound T_4 is only four-to-five times that of free triiodothyronine (FT_3) [2].

Although the majority of the circulating thyroid hormones are protein-bound, only the free forms are readily taken up by cells in the peripheral tissues. Binding proteins act as a buffer to maintain the free fractions within very narrow limits. Bound T_3 and T_4 then serve as a "reservoir" of functional hormone. Changes in concentration of thyroid hormone-binding proteins, which will be discussed later in this chapter, may alter the total but not free hormone concentrations. Increased concentrations of thyroid hormone-binding proteins in the blood cause more T_3 and T_4 to be bound. This, in turn, reduces the amount of "free" thyroid hormone in the circulation. Since FT_3 and free thyroxine (FT_4) concentrations are tightly regulated via the HPTA, this triggers TSH secretion by the pituitary, which stimulates the release of thyroid hormones from the thyroid. This leads to a redistribution of the thyroid hormone between the bound versus unbound states. The end result is a new steady state where the total (i.e., bound plus unbound) thyroid hormone concentration is increased, but concentration of the free form remains the same. This has significant implications on the testing and interpretation of thyroid hormone measurements, and will be discussed more later.

Thyroid hormone transport, conversion, and mechanism of action

Given that T_3 is considered the more biologically active hormone, the main downstream effects of thyroid hormones require the following steps: (1) transportation of T_4 and T_3 from the thyroid to the tissue; and (2) conversion of T_4 to T_3.

Thyroxine and triiodothyronine transport to tissues

In circulation, over 99% of T_3 and T_4 are bound to plasma proteins, namely thyroxine-binding globulin (TBG), transthyretin (TTR, also known as prealbumin), and albumin. The binding characteristics of these proteins are summarized in Table 40.3. In general, T_4 is more tightly bound than T_3 to these plasma binding proteins. Of note, TBG is found in the lowest concentration but has the highest affinity for and binds the greatest percentage of thyroid hormones. Conversely, albumin has the lowest affinity but is present at a significantly higher concentration in the circulation compared with the other thyroid hormone-binding proteins. Several factors can affect the concentration of the thyroid hormone-binding proteins (Table 40.4). As described earlier, changes in thyroid hormone-binding protein concentration can cause a redistribution of free versus bound thyroid hormone. This can lead to a new steady state, where the total thyroid hormone concentration is increased, but concentration of the free/unbound thyroid hormone remains the same.

Conversion of thyroxine into triiodothyronine

Thyroid hormone conversion is a finely tuned process. This is important because different thyroid hormones have varying degrees of biological activity, e.g., T_3 being most biologically active. Regulation of the conversion process can therefore directly influence the overall impact thyroid hormones have on the body. At the peripheral tissue, T_3 and T_4 enter from the circulation via specific thyroid hormone transporters. T_4 can be converted to T_3, the more biologically active hormone, or rT_3, a biologically inactive form of the hormone (Fig. 40.4). Both conversions are catalyzed by iodothyronine deiodinase, which removes an iodine atom from T_4 to generate T_3 or rT_3.

Deiodination can occur at different tissues and to varying degrees, but the liver and kidneys are considered primary sites. Peripheral deiodination of T_4 to T_3 is responsible for producing approximately 80% of the circulating T_3. The remaining circulating T_3 is derived directly from the thyroid gland. The deiodination process controls the balance between biologically active versus inactive thyroid hormone. Under normal physiological condition, peripheral deiodination generates roughly equal amounts of T_3 and rT_3. Approximately 40% and 45% of T_4 are deiodinated into T_3 and rT_3, respectively [3].

Reverse triiodothyronine

T_3 and rT_3 differ only by the position upon which iodine is attached to the tyrosine ring (Fig. 40.3C). Physiologically, T_3 can increase the metabolic rate of most tissues, whereas rT_3 is metabolically inactive.

TABLE 40.3 Characteristics of thyroid hormone-binding proteins.

Plasma binding protein	Thyroxine-binding globulin	Transthyretin	Albumin
Percent of hormone bound	~75% of total T_4	~20% of total T_4	~5% of total T_4
	~75% of total T_3	<5% of total T_3	~20% of total T_3
Concentration in blood	~0.37 μM *(lowest)*	~4.6 μM *(intermediate)*	~590 μM *(highest)*
Binding affinity, *that is, how tightly it binds to thyroid hormones*	$K_a = 10^{10}$ *(highest; higher affinity for T_4 than T_3)*	$K_a = 10^7$ *(intermediate)*	$K_a = 10^5$ *(lowest)*
Binding avidity, *that is, number of hormone-binding sites per protein*	One	Two	Multiple

Ka, Association constant; *T_4*, thyroxine; *T_3*, triiodothyronine.

TABLE 40.4 Examples of factors affecting thyroid hormone-binding protein concentrations.

	Effect on concentration	Explanation
Condition / disease		
• Cirrhosis	↓	Decreased hepatic production of binding proteins
• Nephrotic syndrome	↓	Increased renal clearance of binding proteins
• Malnutrition	↓	Decreased production of binding proteins
• Pregnancy	↑	Estrogen stimulation of TBG production by the liver
• Acute hepatitis	↑	Increased TBG synthesis during acute-phase response
• Acute intermittent porphyria	↑	Mechanisms not clearly established
Drugs		
• Estrogen	↑	Stimulation of TBG production by the liver
• Androgens	↓	Mechanisms not clearly established
• Glucocorticoids	↓	Suppression of TBG synthesis
Others		
• Genetics	↑ or ↓	Genetic variations can lead to ↑ or ↓ production of binding proteins

↑, Increased; ↓, decreased; *TBG*, thyroxine-binding globulin.

Although T_3 and rT_3 are produced at approximately the same rate within peripheral tissues, medications, chronic illness (e.g., chronic liver or kidney disease), medical/surgical stress, or starvation can alter this balance. In the latter state, a lower metabolic rate is physiologically desirable to decrease energy expenditure. As such, increased conversion of T_4 to rT_3 has often been described as a compensatory mechanism by the body to aid survival. In clinical scenarios, such as the ones listed above (also see "Nonthyroidal illness" section), increased conversion of T_4 into rT_3 may ultimately cause a decrease in the T_3/rT_3 ratio [4]. The T_3/rT_3 ratio may provide more clinical information than rT_3 measurements alone. Overall, measurement of rT_3 is considered controversial due to large practice variation in test ordering patterns and little evidence to support frequent ordering of this test [5].

Mechanism of action

Thyroid hormones act on virtually every cell type in the body, with T_3 being the most biologically active form. Until 1970, it was assumed that the lipophilic nature of thyroid hormones allowed them to enter target cells via simple (passive) diffusion. It is now understood that T_4 and T_3 enter the target cells of most organs from the circulation via carrier-mediated transporters, and they require binding to specific intracellular

receptors in order to exert their actions within the cell [6]. Inside the target cells, T_3 exerts most of its actions through regulation of gene transcription. Briefly, T_3 can: (1) bind to thyroid hormone nuclear receptors inside the nucleus; (2) form a complex that can bind to specific DNA sequences; and (3) upregulate or downregulate the transcription of specific target genes. This subsequently affects the amount of the target mRNAs transcribed and proteins synthesized.

The above process allows the thyroid hormones to have a myriad of effects on different cell types, target tissues, and organ systems (Table 40.1). Tissue responsiveness to thyroid hormone can be dependent upon the type and distribution of nuclear receptors within the tissue. The mechanism of thyroid hormone bioactivity may be regulated via processes, such as T_3 conversion rates, the distribution of transport carriers on various cell types, and the type of nuclear receptors located within the cell. These continue to be areas of active research.

Another critical component of thyroid hormone function is reliant upon the relationship between serum concentrations of TSH and FT_4. TSH production by the pituitary is tightly regulated to maintain optimal thyroid hormone concentrations in the blood, predominantly via the negative feedback loop of the HPTA. To accomplish this, there is an inverse log−linear relationship between changes in circulating thyroid hormone and production of TSH, that is, a two-fold change in FT_4 concentration in the blood results in a 100-fold (inverse) change in TSH (Fig. 40.5). As such, measurement of serum TSH concentration is highly sensitive to changes in serum thyroid hormone concentration, and thus diagnostically useful as the first indication of disease. This is particularly important in scenarios where TSH may be abnormal despite normal serum FT_4 (referred to as subclinical disease, see the next section).

Clinical assessment of thyroid function

Overview

In general, thyroid hormones stimulate physiological functions, e.g., cardiac output and heat production. Therefore too much thyroid hormone production (hyperthyroidism) can lead to signs and symptoms of a hypermetabolic state, while thyroid hormone deficiency (hypothyroidism) can result in the opposite presentation. The asymptomatic state with normal thyroid gland function is referred to as euthyroid. An overview of common causes and symptoms of hyperthyroidism and hypothyroidism is summarized in Table 40.5.

Thyroid disease may result from endogenous [e.g., autoantibodies (autoAbs)] or exogenous (e.g., drugs) causes, and present with either overt or subclinical ("mild") conditions. In overt thyroid conditions, signs and symptoms are often present, along with abnormal thyroid laboratory results (Fig. 40.6). In subclinical cases, patients may present with no or minimal clinical symptoms, along with normal to slightly abnormal laboratory results.

Thyroid diseases may also be classified as primary, secondary, or tertiary, depending on what organ the dysfunction originated in. In primary thyroid conditions, the pathology is localized to the thyroid. Oversecretion or undersecretion of hormone by the thyroid will consequently affect the negative feedback loop and overall TSH and TRH production. In secondary and tertiary (also jointly referred to as "central") conditions, the problem originates *outside* the thyroid, typically within the pituitary and hypothalamus, respectively. In these cases, production of TSH (and, less frequently, TRH) is directly affected by pituitary or hypothalamic dysfunction, and downstream production of thyroid hormone will be affected. Primary thyroid disease is much more common

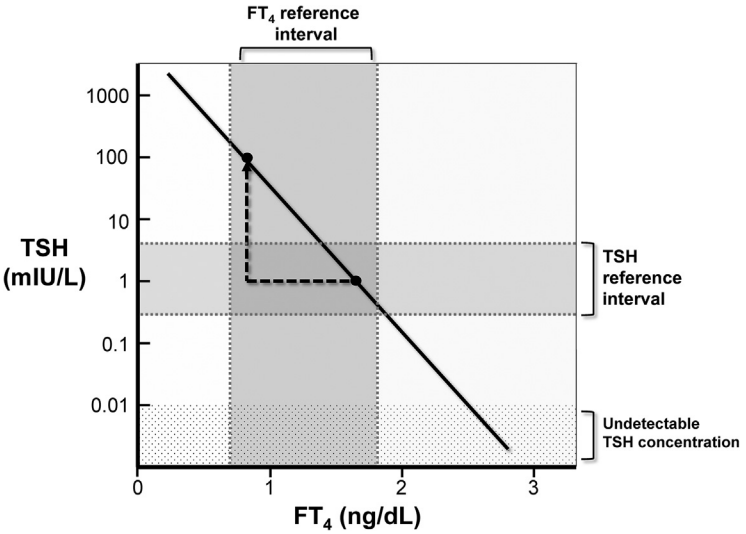


FIGURE 40.5 Graphical representation of the inverse log−linear relationship between serum thyroid-stimulating hormone and free thyroxine concentrations. The logarithmic relationship dictates that small fluctuations in free thyroxine concentrations will be reflected in large differences in thyroid-stimulating hormone values. This relationship explains the utility of thyroid-stimulating hormone as the preferred initial screen for thyroid dysfunction. Dashed arrow: two-fold change of free thyroxine (from 1.6 to 0.8 ng/dL) corresponding to a 100-fold change in thyroid-stimulating hormone (from 1 to 100 mIU/L).

TABLE 40.5 Examples of causes and symptoms associated with hyperthyroid and hypothyroid disease states.

	Hyperthyroidism	Hypothyroidism
Causes	Endogenous	Endogenous

	• Autoimmune thyroid disease
	  ◦ Graves' disease
	• Toxic nodules / toxic multinodular goiter
	• Solitary toxic adenoma
	• Gestational thyrotoxicosis
	• Pituitary tumor
	• Thyroid neoplasm
		• Autoimmune thyroid disease
		  ◦ Hashimoto's thyroiditis
		• Congenital, for example, inborn errors of thyroid hormone biosynthesis
		• Idiopathic atrophic hypothyroidism
		• Pituitary tumor

	Exogenous	Exogenous
	• Excess T_4 or T_3 intake • Exogenous iodine or iodine-containing drugs	• Iodine deficiency • Iatrogenic therapy-related, such as radioiodine (I^{131}), postsurgery, and antithyroid drugs
Signs and symptoms	• Increased metabolism • Weight loss • Increased cardiac output • Heat intolerance • Hyperventilation • Goiter • Tremor • Warm and moist skin • Eyelid retraction	• Decreased metabolism • Weight gain • Decreased cardiac output • Cold intolerance • Hypoventilation • Goiter • Slow muscle relaxation • Lethargy

T_3, Triiodothyronine; *T_4*, thyroxine.

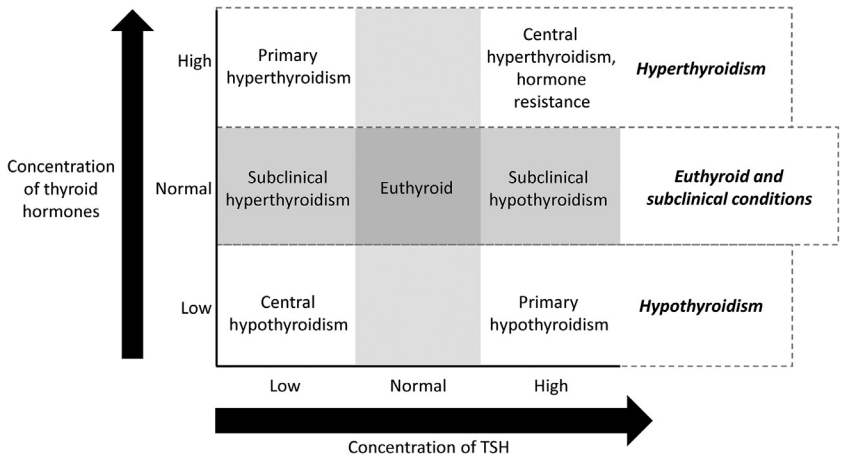


FIGURE 40.6 Relative concentrations of thyroid hormones (triiodothyronine and thyroxine) and thyroid-stimulating hormone in different hyperthyroid and hypothyroid states. Primary and central diseases of the thyroid can be differentiated by thyroid hormone and thyroid-stimulating hormone measurements. The subclinical state is often asymptomatic in presentation, with thyroid-stimulating hormone values outside of the reference interval while free thyroxine remains normal.

than central conditions; this chapter will focus on primary and secondary hyperthyroidism and hypothyroidism.

From the laboratory's perspective, investigation of hyperthyroidism or hypothyroidism generally involves an analysis of hormones associated with the HPTA, namely TSH, T_4, FT_4, T_3, and/or FT_3. Table 40.6 compares the expected changes in TSH, T_4, and T_3 concentrations in primary and secondary disorders. A general overview of a typical testing algorithm for thyroid function evaluation is also shown in Fig. 40.7.

Autoantibodies

Autoimmune thyroid diseases (AITDs) are the most common autoimmune diseases and include both hyperthyroid and hypothyroid conditions (Graves' disease and Hashimoto's thyroiditis, respectively). While the presence of autoAbs—or antibodies (Abs) that bind to self--antigens—in thyroid disease will be discussed further within their respective sections, their central role in AITD warrants a more detailed introduction. Thyroid autoAbs

TABLE 40.6 Comparison of laboratory results between primary and secondary hyperthyroid and hypothyroid disease states.

	Primary hyperthyroidism	Secondary hyperthyroidism	Primary hypothyroidism	Secondary hypothyroidism
Disease	Graves' disease	Pituitary adenoma[a]	Hashimoto's thyroiditis	Pituitary adenoma[a]
Thyroid hormone	↑①	↑②	↓①	↓②
TSH	↓②	↑①	↑②	↓①

↑, Increased concentration; ↓, decreased concentration; ①, initial change; ②, secondary response; *TSH*, thyroid-stimulating hormone.
[a]*Both secondary hyperthyroidism and hypothyroidism may be caused by pituitary adenoma. However, pituitary adenoma is more commonly associated with secondary hypothyroidism, as thyroid-stimulating hormone-producing pituitary adenomas are rare.*

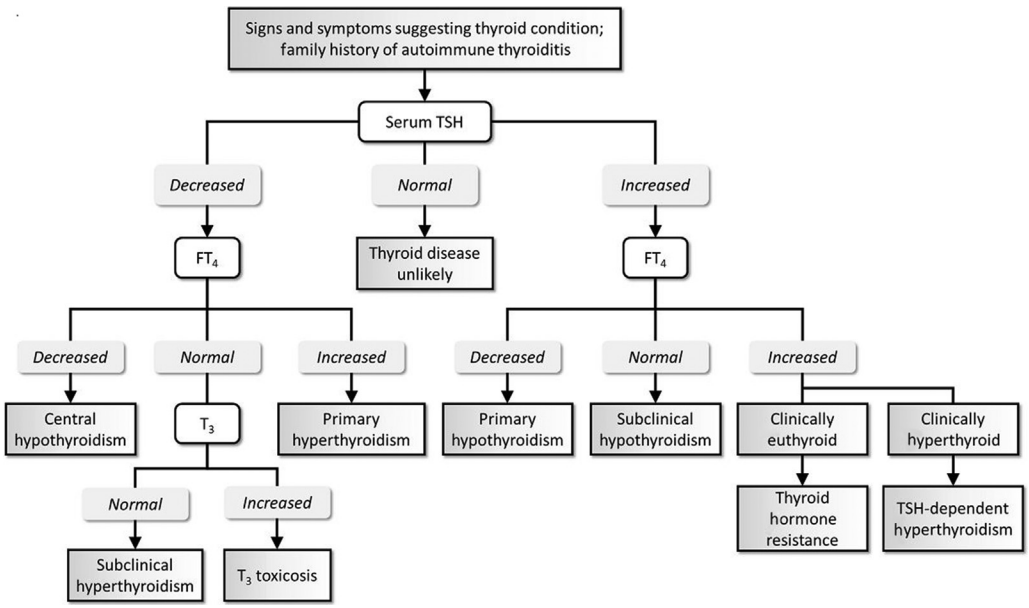


FIGURE 40.7 Example of a typical thyroid disorder diagnostic algorithm. Thyroid-stimulating hormone is central to common testing algorithms, as the best test to screen for initial disease. Thyroid autoantibody testing may be considered in some conditions.

are not usually necessary for diagnosis, but may prove useful in correlation with other thyroid hormone measurements and the clinical presentation of the patient. AutoAbs can target different protein components of the thyroid (see Table 40.7). These include proteins that are critical for the thyroid to receive regulatory signals from the pituitary gland [e.g., the thyroid-stimulating hormone receptor (TSHR)], as well as those involved in the synthesis (e.g., TPO) and transport (e.g., Tg) of thyroid hormones.

Thyroid-stimulating hormone receptor antibodies (TSHRAbs, or simply TRAbs) is a term that encompasses two types of autoAbs: (1) those that stimulate; and (2) those that block the function of the TSHR (neutral autoAbs have also been described, but their function is not well defined). Thyroid-*stimulating* immunoglobulins (TSIs) bind the TSHR and induce the same cAMP signaling pathways that TSH is capable of inducing, resulting in: (1) proliferation of thyroid cells; and (2) synthesis and secretion of thyroid hormones by the follicular cells. This increased production of thyroid hormone leads to the hypermetabolic symptoms associated with Graves' disease, a hyperthyroid condition. Conversely, TSHR-*blocking* Abs, also known as thyrotropin-binding inhibitory immunoglobulins (TBIIs), prevent TSH binding and the subsequent signal cascade required to produce thyroid hormone, leading to the hypothyroid state. Somewhat confusingly, TSI and TBII can actually have overlapping epitope specificities on the TSHR, and both may be found within the same patient. TRAbs are not routinely detected in the general population. The specifics of TRAb and TSI assays will be detailed in "Laboratory measurements of thyroid function."

AutoAbs may also be formed against TPO [thyroid peroxidase antibody (TPO Ab)] and Tg [thyroglobulin antibody (TgAb)], and, when present, they are commonly

TABLE 40.7 Common thyroid-related autoantibodies found in autoimmune thyroid disease states.

Target of AutoAb	AutoAb	Typical findings in different disease states		
		Healthy population	Graves' disease	Hashimoto's thyroiditis
TSHR	TSI	--	++	--
	TRAb	--	-- / +	-- / +
TPO	TPO Ab	~10%	+	++
Tg	TgAb	~10%	+	+

--, Absent; +, less commonly present; ++, commonly present.
Tg, Thyroglobulin; *TgAb*, thyroglobulin antibody; *TPO Ab*, thyroid peroxidase antibody; *TPO*, thyroid peroxidase; *TRAb*, thyroid-stimulating hormone receptor antibody; *TSHR*, thyroid-stimulating hormone receptor; *TSI*, thyroid-stimulating immunoglobulin.

found within the same individual. These Abs tend to be more common in women than in men, and their prevalence increases with age. TPO Ab and TgAb have been identified in approximately 10% of healthy individuals [7,8]. However, due to differences in assay sensitivity and specificity, and therefore reference intervals, reports vary as to whether TPO Ab and TgAb are truly present in the healthy, euthyroid population. With respect to clinical utility, whether TPO Ab presence may act as a predictor of future hypothyroidism has not been confirmed and is likely dependent upon the sensitivity of the assay used. TgAb may be present in individuals with thyroid cancer, but have not been definitely associated with disease progression. TgAb presence is known to interfere with the measurement of Tg; therefore concurrent TgAb measurements are recommended to ensure the validity of Tg results (see "Laboratory measurements of thyroid function" section). Serial measurement of these autoAbs is not necessary or recommended for routinely monitoring disease progression or response to treatment [9].

Hyperthyroidism (thyrotoxicosis)

Hyperthyroidism is defined as the excess production of thyroid hormones. The most common clinical features include weight loss, heat intolerance, warm and moist skin, and tremor and goiter (swelling at the front of the neck due to enlargement of the thyroid gland) (Table 40.5). Of note, the presence of goiter can result from both hyperthyroid and hypothyroid conditions. Another hallmark symptom that may be present is bilateral eyelid retraction (lag), which is the result of exophthalmos, an abnormal protrusion of the eyeball(s). The term thyrotoxicosis refers to the clinical effects of excess thyroid hormone, that is, increased serum FT_4 and/or FT_3, regardless of the cause.

Graves' disease, toxic nodules (or toxic multinodular goiter), and solitary toxic adenoma together account for more than 90% of hyperthyroid cases in the United States. Hyperthyroidism can also, somewhat paradoxically, be observed in the early stages of autoimmune hypothyroidism (Hashimoto's thyroiditis). In this case, immune-mediated destruction of the thyroid at disease onset can release stored thyroid hormone into circulation, appearing as hyperthyroidism (referred to as Hashitoxicosis). Gain-of-function mutations in the TSHR have also been described. Central hyperthyroidism is rare, but may be caused by pituitary adenoma and mutations, resulting in thyroid hormone resistance. Exogenous causes of hyperthyroidism often involve excess thyroid hormone intake or iodine-containing drugs.

Laboratory findings

TSH may aid in differentiating primary versus secondary hyperthyroidism. In general, serum TSH concentration will be low in primary hyperthyroidism and high in secondary hyperthyroidism (see Table 40.6). Other laboratory results consistent with hyperthyroidism include increased serum FT_4, total T_4, and total T_3. In 10%–15% of hyperthyroid patients, T_3 concentration may be increased, while T_4 remains normal. This is known as T_3 toxicosis, and is often observed during the progression of Graves' disease from the subclinical to the overt state.

Graves' disease

Graves' disease is the most common cause of hyperthyroidism. It accounts for 60%–80% of hyperthyroidism cases, and is more common in women than men (5:1 ratio). It also most commonly presents in individuals between 40 and 60 years of age. Graves' disease is an AITD that results in overproduction of thyroid hormone and frequently involves TRAb or, more specifically, TSI. These autoAbs chronically stimulate the TSHR to produce thyroid hormone. Perhaps counterintuitively, TBII and neutral Abs may also be present in these patients, reflecting the heterogeneous nature of the Abs produced in AITD. Although multiple Ab types may be present in Graves' disease, the clinical effects of hyperthyroidism result when TSI activity dominates the activity of any blocking autoAbs that may be present.

Excess thyroid hormone elicits the myriad of symptoms observed, including weight loss, warm and moist skin, and heat intolerance. The classic triad of clinical signs in Graves' disease includes goiter, exophthalmos, and pretibial myxedema (characteristic waxy skin lesions). Orbitopathy is observed in approximately 30% of patients diagnosed with Graves' disease. It has been proposed that TSI binding to TSHR expressed in retro-orbital tissue may be the mechanism responsible for exophthalmos observed in this disease.

Laboratory findings

The typical laboratory results observed in cases of Graves' disease can be found in Table 40.8. Along with the clinical signs described above, Graves' disease is diagnosed by classic hyperthyroid laboratory findings: low TSH, elevated FT_4, and/or T_3. While TSIs are present in 90% of diagnosed individuals and their presence is considered a classic finding in Graves' disease, TSI measurements are often not necessary for definitive diagnosis of the hyperthyroid patient with orbitopathy [10]. TSI and/or TRAb may be helpful in cases with nonclassical presentation, mild symptoms, or no identifiable goiter. They may also be useful in discriminating between AITD and nonimmune causes of thyrotoxicosis, such as toxic nodular goiter. Other autoAbs such as TPO Ab and TgAb may also be present in approximately 80% or 40%–70%, respectively, of patients diagnosed with Graves' disease (Table 40.7). However, these autoAbs are not exclusive to Graves' disease and are more commonly found in Hashimoto's thyroiditis. Radioactive iodine uptake (RAIU, a measure of iodine usage by the thyroid) is commonly elevated in patients with Graves' disease, indicating hyperactivity of the thyroid.

Nodule/multinodular goiter

Thyroid nodules are an abnormal growth of thyroid cells within the thyroid gland. The term thyroid nodule includes both nodular and multinodular goiters. These nodules, sometimes referred to as "toxic" nodules or "toxic" multinodular goiters, are a relatively common cause of hyperthyroidism, with a prevalence of 5% in women and 1% in men in iodine-replete countries [11]. The likelihood of developing thyroid nodules increases with age. Thyroid nodules are typically benign and originate in the colloid. The nodules can develop as colloid benign nodules, cysts, follicular adenomas (i.e., toxic adenoma; "tumor") or less frequently as carcinoma (7%–15%) [11,12]. Nodules may also result from gain-of-function mutations of the TSHR, which cause hyperactivity of the thyroid tissue.

Laboratory findings

Laboratory results will be similar to those seen in Graves' disease, including decreased TSH and increased thyroid hormones (FT_4, total T_4, and T_3). However, unlike Graves' disease, antithyroid Abs are typically absent and can be used to distinguish between the two disease states. Furthermore, despite the growth of one or more nodules, the thyroid gland is generally not diffusely enlarged, as observed in Graves' disease.

In cases where thyroid ultrasound reveals a high-risk nodule, a fine needle aspirate (FNA) may be ordered for further cytological assessment of malignancy. If there is a family history of thyroid malignancy, genetic testing may also be useful (see "Thyroid cancer" section).

RAIU may be increased in patients with toxic nodules or toxic multinodular goiter. However, thyroid scan results will reveal a heterogeneous or "patchy" pattern, rather than the homogeneous pattern observed in Graves' disease.

TABLE 40.8 Comparison of laboratory findings with different thyroid conditions.

Thyroid condition	TSH	Total T4	FT_4	Total T_3 or FT_3	rT_3
Overt primary hyperthyroidism					
Graves' disease	↓	↑	↑	↑	n/u
Central hyperthyroidism	N or ↑	↑	↑	↑	n/u
Subclinical hyperthyroidism	↓	N	N	N	n/u
Nonthyroidal illness	N or ↓	N or ↓	N or ↓	↓	↑
Subclinical hypothyroidism	↑	N	N	n/u	n/u
Central hypothyroidism	N or ↓	↓	↓	↓	n/u
Overt primary hypothyroidism					
Hashimoto's thyroiditis	↑	↓ or N	↓ or N	↓ or N	n/u

↑, Increased; ↓, decreased; *FT_3*, free triiodothyronine; *FT_4*, free thyroxine; *N*, normal; *n/u*, not useful (results are typically normal); *rT_3*, reverse triiodothyronine; *T_3*, triiodothyronine; *T_4*, thyroxine; *TSH*, thyroid-stimulating hormone.

Hypothyroidism

Although statistics vary in the literature, hypothyroidism is more commonly found in the population than hyperthyroidism; approximately 4% of the population may be considered hypothyroid, while approximately 1% is considered hyperthyroid [7]. Hypothyroidism is defined as deficient production of, or impaired response to, thyroid hormones. Symptoms associated with hypothyroidism include weight gain, cold intolerance, lethargy, and goiter (Table 40.5).

Hashimoto's thyroiditis is the most common endogenous cause of primary hypothyroidism in developed countries. Less common causes include idiopathic atrophic hypothyroidism and congenital diseases. Pituitary tumors and adenomas are fairly rare, and their development may lead to either secondary hyperthyroidism or hypothyroidism; the latter is more common. Central hypothyroidism may also result from hemorrhage or trauma, as well as rare genetic mutations.

Exogenous causes of hypothyroidism are often iatrogenic and associated with hyperthyroidism therapy (e.g., radiation-induced hypothyroidism, antithyroid drugs) or surgical procedures (e.g., removal of thyroid gland). Drug-induced hypothyroidism may also occur, most notably from agents such as amiodarone and glucocorticoids. Iodine deficiency is the most common cause of hypothyroidism worldwide, although it is exceedingly rare in developed countries with adequate iodine fortification.

Laboratory findings

Similar to the case of hyperthyroidism, TSH is also a useful marker to differentiate primary versus secondary hypothyroidism. In general, TSH concentrations will be elevated in primary hypothyroidism and decreased in secondary hypothyroidism (Table 40.6). Other laboratory results consistent with hypothyroidism include decreased serum FT_4 and total T_4. Analysis of T_3 is not generally considered a useful test in suspected hypothyroidism as T_3 may remain within normal limits for much of the disease course.

Thyroid replacement therapy (usually with T_4, more rarely with T_3) is used to treat hypothyroidism. T_4 suppresses the pituitary secretion of TSH, which can be measured to monitor response to therapy. Ideally, hormone replacement will return TSH to within the reference interval. If the plasma TSH concentration is not within the reference interval—high TSH suggests inadequate treatment, whereas low TSH may suggest excessive replacement—then FT_4 and FT_3 (in cases where T_3 supplementation is used) should also be measured. This is in part, because the change in plasma TSH concentration can be delayed behind the increases in FT_4 and/or FT_3 from the replacement therapy.

Thyroid autoAbs are often present in hypothyroidism, and their detection can help in the differential diagnosis of difficult or nonclassical hypothyroid cases. Almost 90% of hypothyroid individuals will have thyroid autoAbs in their circulation, predominantly TPO Ab and TgAb. In contrast to hyperthyroidism, RAIU is not as frequently performed in hypothyroidism, and is often not necessary for the diagnosis of specific hypothyroid conditions, such as Hashimoto's thyroiditis.

Hashimoto's thyroiditis

Hashimoto's thyroiditis, or chronic autoimmune thyroiditis, is the most common cause of hypothyroidism in iodine-sufficient countries. It is also currently considered the most common autoimmune disease and most common endocrine disorder [13]. It is a condition characterized by autoimmune destruction of the thyroid gland. Like Graves' disease, autoAbs are also commonly found in patients with Hashimoto's thyroiditis, primarily TPO Ab. Although this condition may occur at any age in either of the sexes, it is significantly more common in females (7:1), particularly in women between 40 and 50 years of age. Gradual loss of thyroid function proceeds to permanent hypothyroidism in a small percentage of cases.

Laboratory findings

The typical laboratory results observed in Hashimoto's thyroiditis are shown in Table 40.8. Depending on the stage of the disease, the FT_4, total T_4, total T_3, and FT_3 may be decreased or remain normal. However, as is common in thyroid disease, TSH tends to be the first analyte to become significantly increased. During the initial subclinical stage, TSH is elevated, while T_4 and T_3 remain normal. As the disease progresses, FT_4 concentrations will begin to decrease, followed by T_3. In prolonged disease, both FT_4 and T_3 can be abnormally low.

In Hashimoto's thyroiditis, TPO Ab and TgAb are present in approximately $>90\%$ and $60\%-80\%$ of diagnosed individuals, respectively [13]. Thyroid autoAbs do not need to be serially measured in patients with overt hypothyroidism; however, these Abs may indicate progression to overt disease in individuals with subclinical, or mild, hypothyroidism [9]. The presence of TgAb provides very little additional diagnostic information over Abs, such as TPO Ab, which are present at higher frequency in Hashimoto's thyroiditis.

Other causes

Hypothyroidism can result from various other causes, including: (1) iodine deficiency; (2) iodine excess; and (3) congenital hypothyroidism.

Iodine deficiency is a rare cause of hypothyroidism in North America due to the availability of iodine in the diet; however, it is the leading cause globally, affecting approximately 2-billion people worldwide. Interestingly, iodine *excess* can also cause hypothyroidism through the Wolff–Chaikoff effect. The Wolff–Chaikoff effect involves suppression of I_2 organification in the presence of excess I^-. This, in turn, suppresses thyroid hormone synthesis in a paradoxical effort to avoid excess hormone production when surplus iodine is present [14].

Congenital hypothyroidism is one of the most common preventable causes of intellectual disability [15]. It is defined as the presence of thyroid hormone deficiency at birth and occurs in approximately 1:4000 of newborns. Roughly 85% of cases are sporadic and due to abnormal development of the thyroid, with the remainder due to hereditary causes. Congenital hypothyroidism may also be transient. The most common cause of transient congenital hypothyroidism worldwide remains iodine deficiency. In cases where the mother has AITD, maternal thyroid autoAbs may cross the placenta and lead to congenital hypothyroidism in the newborn. This cause is uncommon, and the effects are usually transient.

Laboratory findings

In iodine deficiency, TSH may be normal to elevated, while the T_4 may be normal to decreased, and T_3 is normal or mildly elevated. The mild elevation of T_3 is thought to be in part due to preferential secretion by the thyroid and/or increased conversion of T_4 into T_3 in peripheral tissues. This is an important feature, as T_3 is significantly more potent than T_4 but requires only 75% of the amount of the iodine for hormone synthesis. In contrast, in severe iodine deficiency, an increased serum TSH level is more likely to be accompanied by a pronounced decrease in both T_4 and T_3. Iodine excess, traditionally due to pharmaceutical sources such as radiocontrast dyes and amiodarone, has a different presentation depending upon the dosage and/or duration. In acute iodine excess, TSH may be increased, but T_4 and T_3 are only mildly increased. In chronic iodine excess, the patient may be clinically euthyroid, and serum TSH, T_4, and T_3 may all appear to be normal due to presumed "escape" from the Wolff–Chaikoff effect and continued organification of iodide. Chronic iodine excess may be associated with increased incidence of goiter and either hyperthyroidism or hypothyroidism, depending on the clinical scenario.

In congenital hypothyroidism, serum TSH concentration may be elevated, while the thyroid hormone concentrations (e.g., total or FT_4) are low or low-normal. Depending on the laboratory findings, congenital hypothyroidism can also be further classified as subclinical (high TSH and normal FT_4), primary (high TSH and low FT_4), or central (low or normal TSH and low FT_4). Most screening programs for congenital hypothyroidism involve TSH determination, supported by T_4 measurements.

Nonthyroidal illness

It is possible to have abnormal thyroid test results in the absence of thyroid disease. This typically occurs in severely ill, hospitalized, or nutritionally deprived patients, and has also been referred to as "sick euthyroid syndrome," "euthyroid sick syndrome," or "low T_3 syndrome." Hospitalized patients often have altered protein production, circulating factors that inhibit protein binding, and may have received medications that affect thyroid function. Due to this, it is widely accepted that thyroid function should not be assessed in severely ill patients, unless there is a strong indication for thyroid dysfunction. Additional causes of nonthyroidal illness (NTI) include trauma, malignancy, and inflammatory disease.

While the pathogenesis of this condition has yet to be fully elucidated, it is presumed that a lower metabolic rate is physiologically desirable during states such as chronic illness and starvation, as it decreases the overall energy expenditure. Peripheral conversion of T_4 to T_3 therefore declines markedly during starvation and illness. Increased deiodination of T_4 to rT_3, along with concurrent decreases in T_3 production and rT_3 degradation, results in an apparent increase in rT_3 concentrations relative to T_3. Conversion of T_4 to the biologically inert rT_3 has been described as a survival mechanism; however, whether this physiological change is truly an adaptive response remains controversial.

Laboratory findings

Common laboratory findings observed in this population can be found in Table 40.8. Importantly, TSH measurements should not be used alone to assess thyroid function in patients with NTI. Low serum concentrations of T_3, followed by FT_4 and T_4, are commonly found, and TSH may also be low or low-normal. The decrease in serum FT_3 and total T_3 is in part due to decreased conversion of T_4 into T_3 and decreased elimination of rT_3.

Serum T_3 may be particularly useful to measure in hospitalized patients with low serum TSH to help distinguish between hyperthyroidism (elevated T_3) and NTI (low T_3). Although concentrations of rT_3 may be

significantly increased in NTI patients, measurement of rT_3 is not particularly useful. Rarely, it may aid in distinguishing between central hypothyroidism (low rT_3) and NTI (elevated rT_3).

Overall protein concentrations (including thyroid hormone-binding proteins TBG, TTR, and albumin) may be low in severely ill patients, which can affect interpretation of thyroid hormone measurements. For example, T_4 measurements may be low due to TBG deficiency rather than truly impaired synthesis; therefore care must be taken when considering total hormone concentrations in this population.

For the reasons listed above, serum FT_4 concentrations are considered a more reliable marker of thyroid status than total T_4, particularly in patients with NTI. The method used to detect FT_4 is also important, because results may be affected by changes in thyroid hormone-binding capacity (see "Laboratory measurements of thyroid function" section).

Thyroid function and pregnancy

While certainly not a pathological condition, the physiological changes and clinical ramifications of thyroid function during pregnancy warrant their own discussion.

The increased metabolic requirements of the growing fetus cause considerable changes in the thyroid physiology of the mother. During pregnancy, there are increases in iodine requirements for the developing fetal brain, therefore an observed increase in thyroid gland size and thyroid hormone production by the mother. Expected changes in physiology during pregnancy lead to elevations in T_3 and T_4 (due to hCG stimulation of the TSHR), decreased TSH (due to negative feedback following hCG-mediated stimulation of the TSHR), and increases in TBG concentrations (due to estrogen-mediated stimulation of overall protein production).

Graves' disease is the most common cause of autoimmune hyperthyroidism in pregnant women, affecting up to 1% of all pregnancies [16]. More commonly, "gestational thyrotoxicosis" results from hCG's ability to stimulate the TSHR (recall hCG and TSH share the same alpha subunit). The markedly elevated hCG concentrations present within the first and early second trimesters may cause transient hyperthyroidism in 1%–3% of pregnancies [17]. Gestational thyrotoxicosis may also be associated with hyperemesis gravidarum. Rarely, molecular variant forms of hCG or TSHR mutations can also result in gestational thyrotoxicosis.

Overt hypothyroidism affects up to 0.5% of all pregnancies, with 2%–3% of pregnancies affected by subclinical hypothyroidism [18], primarily due to Hashimoto's thyroiditis. T_4 plays a critical role in neurologic development of the fetal brain, and deficiencies have been associated with diminished neuropsychological development and cognitive function in early childhood [19]. The increased demand for thyroid hormone can naturally result in hypothyroidism later in pregnancy, particularly in women that had borderline or limited thyroidal reserve prior to conception.

Some studies have shown the presence of thyroid autoAbs (particularly TPO Ab) during pregnancy to be associated with increased risk of postpartum thyroid disease. However, most current guidelines for thyroid management during pregnancy do not recommend universal TPO Ab screening in this population due to insufficient evidence of this relationship [18,20]. Recently, the American Thyroid Association (ATA) revised their recommendation to include testing for TPO Ab in pregnant women with TSH values of above 2.5 mU/L [21]. It should also be noted that screening of asymptomatic pregnant women for thyroid function is controversial. Universal screening prior to or during pregnancy is not formally recommended by current practice guidelines in light of conflicting evidence of its overall benefit to mother and child [18,20,21]. However, these guidelines do support thyroid function testing for women at increased risk for thyroid dysfunction, particularly hypothyroidism.

Laboratory findings

As mentioned, the healthy pregnant woman will have elevated laboratory values for TBG, T_4, and T_3, and decreased TSH values compared with nonpregnant individuals. T_4 and T_3 values may be increased up to 1.5-fold over nonpregnant women. Because of these considerable differences in thyroid physiology as the gestational period progresses, trimester- and method-specific reference intervals are suggested, most notably for TSH and FT_4 [21]. Additionally, direct measurements [isotope dilution–mass spectrometry following dialysis or ultrafiltration (UF), see "Laboratory measurements of thyroid function" section] for FT_4 and FT_3 have been recommended in this population due to the abnormalities in binding protein concentrations observed during pregnancy [21,22]. In the absence of this specialized testing, clinical practice guidelines suggest free thyroxine index (FTI; see "Laboratory measurements of thyroid function" section), or total T_4 measurements may be more reliable than immunoassay methods for FT_4 during pregnancy [18,21]. However, this recommendation has been controversial, particularly within the laboratory community [23].

As in nonpregnant individuals, thyroid function during pregnancy is often first evaluated by measuring TSH, with abnormal values further investigated by measuring FT_4 (or comparable measurement). Gestational thyrotoxicosis will result in increased production of T_4 (and rarely,

T_3) and, therefore, low or undetectable TSH. TRAb testing may help distinguish between autoimmune and gestational causes of hyperthyroidism. Hypothyroidism is diagnosed with increased TSH and decreased FT_4 concentrations; however, in all cases, care should be taken to use trimester- and method-specific reference intervals.

Thyroid cancer

Thyroid cancer is the most common cancer of the endocrine system. While the incidence of thyroid cancer has risen by an average of 3.1% annually over the past 10 years, the 5-year survival rate has increased to over 98% [24]. This is likely attributed to an increase in imaging use, which has prompted earlier diagnoses and improved prognoses.

As mentioned earlier, thyroid nodules are an abnormal growth of radiologically distinct cells within the thyroid tissue that are found frequently in the general population and are commonly benign. However, 7%−15% of thyroid nodules are malignant, which dictates the clinical importance of their identification [11,12]. Papillary thyroid carcinoma (PTC) and follicular thyroid carcinoma (FTC) are the two most common types of thyroid cancer. These are differentiated thyroid cancers (DTCs) that derive from malignant follicular cells and together make up 70%−80% and 10%−15% of all thyroid cancers, respectively. Notably, PTC is the most common and least aggressive thyroid malignancy and makes up the majority of new thyroid cancer diagnoses. Medullary thyroid carcinoma (MTC) is responsible for approximately 2% of thyroid cancers and originates from malignancy of the C cells. These are less differentiated and generally more aggressive cancers. Anaplastic thyroid carcinoma is considered the most aggressive undifferentiated thyroid cancer of follicular cells and contributes to less than 1% of all thyroid cancers. The next section will focus on laboratory testing associated with PTC, FTC, and MTC.

Laboratory findings

Suspected thyroid nodules are initially investigated by ultrasound and for functionality (ability to produce thyroid hormone) by measuring TSH concentrations. Low TSH values indicate excess thyroid hormone and are followed up with RAIU testing. "Hot" nodules take up radioactive iodine and indicate hyperfunctioning nodules, which are rarely malignant. Patients with "cold," or nonfunctioning, nodules or elevated TSH values are further investigated by FNA and/or ultrasound.

PTC and FTC are of follicular origin; therefore serum Tg measurement may be useful in the context of thyroid cancer. Of note, current ATA guidelines do not recommend routine measurement (or screening) of serum Tg for the initial evaluation of thyroid nodules [11]. This is based on the finding that serum Tg may be elevated in several other thyroid conditions and lacks both sensitivity and specificity for diagnosis of thyroid cancer. Similarly, routine preoperative measurements of serum Tg or TgAb are not recommended in thyroid cancer patients. However, postoperative serum Tg, for example, measured several weeks after thyroidectomy and prior to radioiodine remnant ablation, can help guide risk stratification and clinical management decisions. Elevated postoperative serum Tg concentrations have been associated with increased risk of DTC recurrence and decreased survival. The risk of persistent disease increases with elevated or progressively higher postoperative Tg concentrations. Serum TgAb measurements should accompany all Tg measurements, due to possible interference (see "Laboratory measurements of thyroid function" section). Rising trends in serial TgAb values have also been suggested as an indicator of disease recurrence and possible surrogate marker for DTC [11,25].

As described earlier, in MTC, the malignancy originates in the C cells, which are largely responsible for secretion of calcitonin. Calcitonin is therefore used as a tumor marker of MTC; however, it may also be elevated in conditions not related to thyroid neoplasms. Current guidelines therefore do not specifically recommend routine measurement of serum calcitonin in patients with thyroid nodules [11]. However, in specific cases where thyroid nodule evaluations via FNA biopsies are inconclusive or suggestive of MTC, calcitonin measurements may be useful to guide diagnosis or treatment approach [26]. Postoperatively, serum calcitonin evaluation may be used to help detect residual disease or recurrence of MTC. Elevated calcitonin after surgery for MTC has been associated with a worse prognosis. Although not a specific marker of the disease, carcinoembryonic antigen (CEA) measurements have also been described for monitoring disease progression in MTC patients postthyroidectomy. However, it is not recommended for use in diagnosis [26]. Calcitonin and CEA measurements should be monitored for doubling times postthyroidectomy, as these rates have been associated with MTC progression [26].

It is worth noting that beyond the circulating thyroid-related biomarkers described (e.g., Tg, TgAb, calcitonin, and CEA), several molecular targets have been identified in different thyroid cancer types. For instance, BRAF gene mutations and RET/PTC (PTC1 or PTC3) gene rearrangements have both been observed in PTC patients. On the other hand, RAS point mutations and PAX8-PPARγ rearrangements have been implicated in FTC, and point mutations of RET have been associated with MTC. Testing panels encompassing these cancer-associated mutations may be performed on FNA samples or excised tissue. Specific molecular profiles, or combinations of

oncogenic mutations, have been associated with worse PTC outcomes [11]. In general, molecular markers are most useful when FNA cytological results are indeterminate. While genetic counseling and testing for MTC should be considered in very specific clinical scenarios, for example, patients with first-degree relatives diagnosed with hereditary MTC, it currently remains unclear whether the use of single or multiple molecular markers is beneficial for managing patients with thyroid nodules in routine clinical practice [11,26]. Several methods are available for detecting the aforementioned alterations in these molecular targets, including IHC, FISH, RT-PCR, and sequencing. The methods available will depend on the mutation of interest, and are described in more detail elsewhere in this text.

Laboratory measurements of thyroid function

Overview

The signs and symptoms of thyroid conditions are generally nonspecific. As such, in addition to medical history and physical examination, laboratory tests for thyroid functions are an important component in diagnosis of the underlying condition. Laboratory assessment of thyroid function typically involves measuring analytes, such as TSH, FT_4, and/or T_3, in the serum or plasma. Accurate measurement is essential and requires adequate sensitivity to detect low concentrations of thyroid hormones. Sufficient analytical specificity is also required to differentiate between structurally similar hormones.

Although published several years ago, many recommendations regarding thyroid testing originated with the Laboratory Medicine Practice Guidelines published by the National Academy of Clinical Biochemistry (NACB, now known as the AACC Academy) [27], which is referenced by many current clinical practice guidelines [9,18,21], including those by the ATA and the American Association of Clinical Endocrinologists (AACE). A commonly followed approach for testing thyroid abnormalities is to start with TSH measurement (Fig. 40.7). The ATA/AACE describe serum TSH as "the single best screening test for primary thyroid dysfunction for the vast majority of outpatient clinical situations" [9]. Recall the inverse log-linear relationship between TSH and FT_4 (see Fig. 40.5). This describes the increased sensitivity of TSH to thyroid hormone changes and its ability to detect abnormalities before FT_4 measurements will. If an individual's TSH measurement is within the defined reference intervals and they are asymptomatic, the patient is considered euthyroid. Due to the diagnostic value of TSH and improved assay measuring ranges accommodating most hyperthyroid and hypothyroid disease states, testing beyond a euthyroid TSH value is most often not required. Values outside of the traditional reference intervals warrant additional testing, usually of FT_4, to investigate hypothyroidism or hyperthyroidism and determine primary versus secondary causes. Testing beyond these points in the algorithm is indicated only in rare situations and is often unnecessary. Proper laboratory test utilization efforts have encouraged the use of testing algorithms, such as this, in an effort to minimize unneeded testing.

Despite reported traceability to reference methods, preparations, or standards, many immunoassays for total and free thyroid hormones and TSH still do not provide comparable results. TSH standardization, in particular, is complicated by its heterogeneity; circulating TSH is a mixture of different glycoforms and therefore not easily standardized against a singular preparation. The International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) formed the Working Group for Standardization of Thyroid Function Tests (WG-STFTs) to address this concern. They have proposed efforts to harmonize TSH and standardize FT_4 assays that are currently underway [28].

It should be noted that screening for thyroid disease in the general population remains controversial, as studies providing direct evidence of the benefits and harms of screening are lacking [9]. While not currently recommended in asymptomatic adults, ATA practice guidelines suggest screening in patients over the age of 60 and in the presence of other abnormal thyroid findings; however, supportive data are limited [9].

Thyroid-stimulating hormone (thyrotropin)

Overview and indications

The inverse log−linear relationship dictates that serum TSH concentrations are highly sensitive to changes in serum thyroid hormone concentrations (recalling a two-fold change in FT_4 leads to a 100-fold change in TSH) and thus often provide the first indication of disease. Modern TSH assays also have improved sensitivity that allows for detection of values expected in both hyperthyroid and hypothyroid states. These two factors have contributed to the current clinical practice guidelines for both hyperthyroidism and hypothyroidism, concluding that serum TSH measurement is the single best screening tool for primary thyroid dysfunction in the majority of outpatient scenarios [9,10]. Measurement of thyroid function in the critically ill and hospitalized populations should still be limited to highly suspect cases.

Methods

Clinical laboratories utilize immunoassays for measuring serum TSH. As noted earlier in this chapter, TSH consists of two subunits: an alpha subunit, which is identical to those found in LH, FSH, and hCG, and an unique beta subunit. Immunoassays for TSH must therefore avoid potential cross-reactivity with LH, FSH, and hCG. TSH concentration is expressed in milli-international units of activity per liter of serum (mIU/L) largely due to historical origins. TSH was originally measured using in vivo and in vitro bioassays, involving animal thyroid tissue and cultured follicular cells. The minimum amount of TSH required to trigger a biological response was defined as the international unit and subsequently adopted as the unit of measure for TSH.

Current methods for measuring TSH concentrations generally employ a two-site "sandwich" heterogeneous immunometric technique. Briefly, these immunoassays involve two Abs—a "capture" Ab and a "signal" Ab. The combination of two Abs confers the specificity required to differentiate between TSH and analytes that share a similar alpha subunit. A variety of detection systems are available, including (electro)chemiluminescent, photometric, fluorescent, or enzymatic methods.

Considerations

Different "generations" of TSH immunoassays have evolved over many years. In short, TSH assays have customarily been classified into different generations based on the limit of quantitation (LOQ) (functional sensitivity; 20% between-run coefficient of variation) achieved. Each generation has exhibited a 10-fold improvement in sensitivity (e.g., 1.0 mIU/L for the first generation, 0.1 mIU/L for the second generation, and 0.01 mIU/L for the third generation assays). Some manufacturers have also claimed 0.001-mIU/L sensitivity for the newer, "fourth" generation assays. Current third generation assays surpass the minimum LOQ of 0.02 mIU/L proposed as the standard of care in the hospitalized setting [27]. It is thought that this level of sensitivity allows for the accurate diagnosis of primary hyperthyroidism (profound TSH suppression), as well as for the differentiation of hyperthyroid from subclinical patients and those with other NTI. Beyond sensitivity, TSH assays also require a sufficiently wide measurement range to accommodate the varied concentrations—up to five orders of magnitude difference—expected between hyperthyroid and hypothyroid conditions.

As with all immunoassays that will be described, potential assay interferences exist and should be considered when interpreting thyroid function tests (reviewed in [29], also see Chapter 12, Immunoassays). These include, but are not limited to, cross-reactivity with similar compounds, the presence of Abs to TSH, heterophile Ab interference with reagents, TSH molecular variants, and drug interferences.

Defining the upper reference limit for TSH is widely debated, with no true consensus. With improvements in TSH assay sensitivity and specificity and better defined reference populations, upper reference limits have declined steadily over time. Given the prevalence and asymptomatic nature of subclinical hypothyroidism, it is reasonable that a small percentage of hypothyroid individuals have been included in historic reference interval determinations. This would result in higher mean values for the upper limits than what would be expected in an entirely healthy population. Previous data suggest that 95% of healthy individuals, without subclinical thyroid disease or the presence of thyroid autoAbs, have TSH values of up to 2.5 mIU/L [27], while current ATA practice guidelines suggest an upper limit of 4.12 mIU/L based on data from the US NHANES III population [7,9]. The value used to define the upper TSH limit of a reference population will affect the number of people investigated for hypothyroidism; thus this remains an important point to be clarified. Currently proposed upper limits vary from 2.5 to 4.5 mIU/L, depending on the professional organization and the population studied [9].

Due to lack of harmonization between assays and differences in epitope recognition, it is recommended that assay- and laboratory-specific reference intervals are determined. Individuals with thyroid autoAbs or other indications of thyroid dysfunction should be omitted from reference populations.

Although most studies have revealed that mean TSH values increase in euthyroid individuals with each decade, with wider variability within age groups, age-specific reference intervals are not recommended within the adult population [7,27,30].

Total thyroxine

Overview and indications

The thyroid gland primarily secretes T_4, which is more protein bound and has a longer half-life than T_3 (Table 40.2). Subsequently, circulating T_4 concentrations are approximately 50−100 times those of total T_3. These higher concentrations, along with the tight regulation of T_4 production by TSH and the HPTA, have resulted in T_4's use as another marker of thyroid function. Total T_4 concentrations are not commonly measured, but may be indicated in the absence of protein-binding abnormalities or to investigate discordant FT_4 results.

Methods

The most common method used to measure total T_4 in the clinical laboratories is immunoassay. Given the structural similarities between T_4, T_3, and rT_3, the anti-T_4 Ab used in these assays must be highly specific for T_4. In addition, because the majority of T_4 in circulation is protein-bound, immunoassays for total T_4 often involve a dissociation step to remove T_4 from tightly binding proteins, such as TBG and TTR. This is commonly achieved using blocking agents, for example, 8-anilino-1-napthalene-sulfonic acid (ANS), and barbital can displace T_4 bound to TBG and TTR, respectively. After this displacement step, both unbound and FT_4 are available, thus allowing for the measurement of "total" T_4. Accurate and precise measurement of total T_4 is therefore dependent upon complete displacement of these T_4−protein complexes.

Unlike TSH assays, where the analyte of interest is a sufficiently large protein molecule compatible with the "sandwich" immunoassay format, T_4 is a considerably smaller molecule. As such, competitive immunoassays are commonly used to measure circulating T_4 concentrations, including chemiluminescent assays, enzyme multiplied immunoassay technique, and cloned enzyme donor immunoassay. The availability and use of pure T_4 standards for assay calibration has allowed for better interassay standardization than with other hormones, but challenges still exist.

Reference methods for total T_4 using isotope dilution−liquid chromatography−mass spectrometry have been described [31] along with other chromatographic methods, but these are not in routine use. As with immunoassay, these methods require liberation of T_4 from binding proteins, often with an acidification step.

Considerations

In clinical practice, total T_4 measurements have largely been replaced by FT_4 assays. Free hormone assays can provide more accurate measurements in cases of protein-binding abnormality, such as pregnancy, severe illness, or in the presence of drugs. In these situations, the total T_4 concentration may change, while the biologically relevant FT_4 remains the same. Historically, total T_4 concentrations have also been used in conjunction with TBG to approximate free hormone concentrations (see "Indirect measurement of free thyroxine: index methods" section).

Free thyroxine

Overview and indications

As described earlier, free or unbound hormone is considered the biologically active form and most directly correlates with the clinical status of the thyroid. As such, the human body puts great effort into maintaining a constant concentration of the free thyroid hormone fraction. This is evident by the fact that alterations in circulating hormone-binding protein concentrations can lead to changes in total T_4 without significant changes in thyroid function. This fact, along with the increasing availability and adoption of FT_4 assays in clinical laboratories, has resulted in FT_4, becoming the preferred test for thyroid diagnostics downstream of initial TSH testing.

Free hormone measurements may utilize one of two approaches: (1) "direct" measurements that physically separate free from protein-bound hormone; or (2) estimated ("indirect") measurements. Estimated measurements of free hormones include mathematical calculations ("index" methods) and immunoassays that approximate the free fraction from the total measurement.

"Direct" measurements have evolved as reference methods for free hormone measurements. Clinically, they are considered most useful in cases where FT_4 and TSH results are discordant with the clinical picture. They are also considered the best measure of free hormone in pregnant patients [21]. "Index" approaches have primarily become obsolete in light of the availability of alternative methods. However, it is still worthwhile to understand how the concentration of FT_4 can be estimated, and why current methods offer a more direct assessment of free thyroid hormone concentration.

Methods

Direct measurement of FT_4

In order to measure FT_4, two challenges need to be overcome: (1) measuring the concentration of FT_4 without disturbing the equilibrium between FT_4 and bound T_4 (most of the T_4 being in the latter state); and (2) obtaining adequate analytical sensitivity to accurately quantify the exceedingly low concentrations of circulating FT_4.

To overcome the first challenge, several approaches have been used, including equilibrium dialysis (ED) and UF. ED involves the separation of bound and FT_4 using a semipermeable membrane. In brief, unbound T_4 are able to freely cross the semipermeable membrane. After reestablishing equilibrium, the dialysate side of the membrane now contains the protein-free, unbound T_4, that is, FT_4. Similarly, in UF, bound ("large") and unbound ("small") hormones are separated by size through a semipermeable membrane during high-speed centrifugation. Following these separation procedures, FT_4 may be directly quantified in the dialysate or ultrafiltrate.

To overcome the second challenge—being able to accurately quantify low concentrations of circulating FT_4—methods including immunoassay and mass

spectrometry (MS) are often employed. Techniques involving further purification of dialysate have also been used. The dialysate is subjected to further extraction methods, such as solid-phase or liquid-phase extractions, to isolate the FT_4 prior to downstream analysis by liquid chromatography–tandem mass spectrometry (LC–MS/MS) [32]. This preprocessing step can further enhance the sensitivity of the analysis, allowing more accurate determination of FT_4 concentration.

Considerations

When direct LC–MS/MS methods for FT_4 have been compared with TSH concentrations, the expected inverse log–linear relationship is evident, with no correlation to TBG or albumin concentrations [33]. In contrast, studies using indirect FT_4 immunoassay measurements show strong correlations with TBG and albumin, and the inverse TSH relationship is absent, suggesting inadequacy of the measurement. While direct measurement of free hormones is considered a superior assessment in the clinical scenarios described, these methods do have limitations. While UF may be a faster procedure compared with ED, it is prone to the effects of temperature changes and absorption of analyte onto the membrane. Both of these can affect the reproducibility of free thyroid hormone recovery. In contrast, ED is susceptible to factors, such as dilution by the dialysis buffer, changes in the dialysis buffer composition, and binding inhibitors included in the dialysis. Nevertheless, both of these methods used in combination with downstream LC–MS/MS analysis have reported comparable results, and ED with isotope dilution–liquid chromatography–tandem mass spectrometry (ID–LC–MS/MS) is the reference procedure proposed by the IFCC WG-STFT for FT_4 [34]. ED with ID–LC–MS/MS is also the recommended measurement for FT_4 in pregnant women [21]. Currently, direct measurements of FT_4 (as well as FT_3) via ED or UF filtration are performed only in specialty laboratories and not routinely available.

Indirect measurement of free thyroxine: index methods

The phrase "index method" refers to a group of calculations that mathematically estimate free hormone concentrations from total hormone and binding protein measurements. Overall, these calculations "normalize" the total hormone concentration with respect to the amount of thyroid hormone-binding protein, primarily TBG. There are several tests that belong to this group (some with multiple names), including T_3 (or T_4)-uptake [T_3U, also referred to as T_3-resin uptake, thyroxine-binding capacity, or simply T-uptake], thyroid hormone-binding ratio, thyroid hormone/TBG ratio, FTI, and T7. Due to their limited use, this chapter will simply highlight two of these tests: T-uptake and FTI.

The T-uptake test approximates the T_4-binding capacity of TBG by determining the number of available T_4-binding sites. Excess labeled T_4 (or, historically, isotopically labeled T_3) is added to saturate any unoccupied binding sites on TBG. It is presumed that the more unoccupied binding sites that are available to bind this labeled hormone, the less T4 hormone that was present in the initial sample. Results are often reported as a percent of the values obtained from a euthyroid population. The T-uptake result alone does not provide much clinical utility; rather, it is more often used as a component of the FTI calculation. FTI is a mathematical approximation of the FT_4 concentration, which is considered a more reliable indicator of thyroid function than total T_4 when alterations in protein-binding concentrations are present. FTI is calculated by multiplying the total T_4 concentration by T-uptake.

By taking the concentration of thyroid hormone-binding proteins into account, this helps compensate for abnormal binding protein concentrations that might be present in patients due to various factors (see Table 40.4). Despite the utility of FTI, alternative methods have been developed to allow direct measurement of FT_4. Currently, direct FT_4 or immunoassay measurements are considered preferable test alternatives to FTI in most clinical situations.

Indirect measurement of free thyroxine: immunoassay

Indirect measurement of FT_4 does not require physical isolation of the unbound hormone fraction prior to analysis. As such, it is the method of choice for automated immunoassay analyzers. Two general approaches have been described: simultaneous one-step and sequential two-step methods. Both methods utilize similar principles; however, the latter is more widely used by automated analyzers.

The "one-step" method is essentially a competitive heterogeneous immunoassay. Briefly, labeled or signal-generating hormone analogs of T_4 (e.g., chemically modified T_4) and anti-T_4 Abs are added to the patient specimen. The FT_4 analog design prohibits binding to serum proteins, but it competes with FT_4 in the patient's specimen for binding sites on the anti-T_4 Abs. These Abs are generally immobilized on a solid support or conjugated to paramagnetic particles to allow capture and separation of the Ab–ligand complex.

By allowing a new equilibrium to be established after the addition of the hormone analog, signal generated by the labeled FT_4 analogs will be inversely proportional to the endogenous FT_4 concentration. In other words, the less endogenous FT_4 that is present in the patient specimen, the more labeled T_4 analog that is able to bind to

the anti-T_4 Abs; upon removal of the Ab−ligand complexes, less signal will be detected.

In the two-step method, the FT_4 in the patient's specimen is first allowed to equilibrate between the endogenous thyroid hormone-binding proteins and the anti-T_4 Abs, which are typically immobilized. After removing the unbound serum, labeled or signal generating T_4 is added, which will bind to the remaining unoccupied sites on the immobilized anti-T_4 Abs. This is followed by a wash step to remove the unbound labeled T_4 before the signal (from the labeled Ab bound to immobilized Ab) is quantified; that is, the more endogenous FT_4 that is present, the less unoccupied binding sites on the immobilized Ab for the labeled T_4 to bind. Ultimately, the signal detected will be inversely proportional to the FT_4 concentration in the patient specimen.

Considerations

Practically, the use of immunoassay for FT_4 measurements can grow out of necessity and the lack of readily available MS-based methods. Indirect FT_4 methods are frequently employed, in part due to the ease of automation. Compared with indirect FT_4 estimates, direct FT_4 values have been shown to correlate better with log-TSH values in patients with abnormal or normal thyroid hormone−protein-binding levels. It has also been demonstrated that LC−MS/MS-based FT_4 measurements, considered the current gold standard, are less affected by albumin or TBG concentrations compared with immunoassay-based methods [33].

In the case of one- and two-step methods, the presence of anti-T_4 Abs could potentially disrupt the equilibrium between bound and unbound thyroid hormones. This, in turn, can affect the determination of the FT_4 concentration. It is therefore essential that only small amounts of total hormone are sequestered by these newly formed complexes (often less than 3% of the total hormone concentration). This amount exceeds the actual free hormone concentration, therefore causing an insignificant disruption of the equilibrium. This allows both methods to provide a fairly accurate approximation of the FT_4 concentration. Nonetheless, measurements using ED and UF followed by LC−MS/MS are still the preferred methods in situations where the patient may have significant changes in thyroid hormone-binding protein concentrations, for example, elevated TBG in pregnant women.

Total triiodothyronine and free triiodothyronine

Overview and indications

T_3 is more biologically active than T_4, but is secreted at lower concentration by the thyroid. Most of the circulating T_3 is generated at the peripheral tissue through conversion of T_4 into T_3. Relative to T_4, T_3 is less protein-bound in the circulation and thus has a shorter half-life (see Table 40.2).

Measurement of FT_3 or total T_3 is indicated only in rare situations. Total T_3 has been used to further investigate samples that have undergone the TSH reflex outlined earlier (i.e., suppressed TSH and FT_4 within reference limits) to distinguish between patients with suspected subclinical hyperthyroidism and T_3 toxicosis. Rarely, T_3 measurements may provide additional clinical information in cases with abnormal T_4 to T_3 conversion at the peripheral tissue, such as in the case of deiodinase enzyme mutation or NTI. Neither total T_3 nor FT_3 is recommended in the evaluation of hypothyroidism, as it is frequently one of the last measures of thyroid function to be affected in this disease state [9].

Methods

Despite lower concentrations, the methods used to quantify total T_3 and FT_3 are analogous to those used for T_4 and FT_4. Immunoassays are available for determining the total and free concentrations of T_3 in the serum. Similar to the total T_4 assays, measurement of total T_3 also requires liberation of bound-T_3 (e.g., via acidification by ANS) before using T_3-specific Ab to determine the total (free + unbound) T_3 concentration. Likewise, the direct measurement of FT_3 can also be accomplished using ED or UF followed by analytical methods, such as immunoassay or MS. The same indirect methods for estimating FT_4 concentrations may also be used to determine FT_3 concentrations.

Considerations

While indications for measurement are rare, when needed, the decision to use total T_3 or FT_3 measurements is often based on preference and assay availability [10]. Many clinicians default to total T_3 measurements, which predated the availability and validation of FT_3 assays. There is currently no guideline universally dictating the use of FT_3 versus total T_3.

Although the methods used to measure T_3 and FT_3 are similar to those used for T_4 and FT_4, T_3 immunoassays tend to have worse interassay precision. This is thought to be attributed to, at least in part, the fact that serum concentrations of total T_3 are 50−100 times lower than those of T_4.

Reverse triiodothyronine

rT_3 is a metabolically inactive form of T_3 generated from T_4 at the peripheral tissue. Its half-life is even shorter

than that of T_3 (hours instead of days), and its concentrations usually follow those of FT_4. The utility of rT_3 assays is limited; it may help support the diagnosis of NTI but often is not necessary. It has also been proposed that the T_3/rT_3 ratio may be a useful prognostic indicator in patients with NTI. In these patients, the TSH may be normal or high-normal, while the total T_3 and T_3/rT_3 ratio are both low. Although not commonly indicated, both radioimmunoassays and LC−MS/MS assays are available for rT_3 measurement, often in specialty laboratories. LC−MS/MS assays allow the simultaneous quantification of rT_3 and total T_3. At present, the use of rT_3 remains controversial [5].

Thyroxine-binding globulin

In cases where thyroid hormone measurements do not align with the clinical picture, or in patients with abnormal total but normal free thyroid hormone concentrations, further investigation into hormone-binding proteins may be warranted. TBG measurements may aid in determining excess (e.g., during pregnancy) or deficiency (e.g., inherited TBG deficiency), which may affect overall hormone measurements. Historically, TBG along with total T_4 provided an assessment of free hormone concentrations; however, this has been largely replaced with modern FT_4 assays. Currently, immunoassays are the most common methods used to measure TBG concentrations. These assays may be competitive or noncompetitive (i.e., "sandwich"), homogenous or heterogeneous. A variety of detection methods may also be used in these assays, for example, fluorescence, chemiluminescence, or turbidimetry. Overall, TBG measurements are not common and not recommended for routine thyroid screening.

Thyroglobulin

Tg is a particularly important marker in the monitoring of thyroid cancer patients (specifically papillary and follicular cancers). As Tg is produced solely by the thyroid, it is a highly specific marker for this tissue. Tg concentrations become undetectable postthyroid ablation therapy or thyroidectomy. Serial monitoring is then initiated, with increased serum Tg concentrations suggestive of cancer recurrence.

Both competitive and noncompetitive immunoassays are commonly used to measure serum Tg. However, the presence of TgAb can interfere with traditional Tg measurements and underestimate its concentration. Given that TgAb has been reported to be detectable in up to 10% of the general population and as high as 25%−30% of thyroid cancer patients [8], measurement of Tg with automated immunometric assays can be problematic in these patients. Considerable variation is observed among these assays.

Historically, radioimmunoassay has been used to provide a more specific measurement of Tg in these situations [35]. More recently, LC−MS/MS-based methods have been developed to circumvent this interference. Often, enzymatic digestion is used to disrupt Tg and TgAb, if they are present. An Ab is then used to capture peptide fragments that are specific to Tg, and these fragments are then analyzed using LC−MS/MS. By determining the relative amount of endogenous Tg-specific peptides and internal standards, absolute quantification of Tg in the patient sample, even in the presence of TgAb, may be achieved [36,37]. Further studies are needed to fully evaluate the use of this newer technology in the TgAb-positive population.

Due to the possibility for interference in immunoassay techniques, NACB and the ATA have both encouraged that patient samples be tested for TgAb whenever serum Tg is measured. With the advent of LC−MS/MS assays that are not susceptible to TgAb interference, some clinical laboratories have adopted a testing algorithm that is initiated with TgAb testing. If TgAb are present, Tg testing will be appropriately reflexed to an LC−MS/MS method. In contrast, if TgAb testing is negative, automated immunoassays will be used to measure Tg. This testing strategy has been shown to encourage proper test utilization and considerable cost savings, given that the immunoassay performance is comparable with LC−MS/MS in the absence of TgAb [38].

Thyroid autoantibodies

A variety of assays are available for the thyroid autoAbs described earlier in this chapter, including TPO Ab, TgAb, TRAb, and TSI. Some common examples of the methods used include ELISA and radioimmunoassay. Automated immunoassays for TPO Ab and TgAb are routinely available; however, considerable variation is observed between different assays, and care should be used when comparing results between platforms or methods.

As mentioned earlier, TSI and TBII may have overlapping specificities, and both may be present in an individual patient (see "Autoantibodies" section). TRAb assays detect the presence of all autoAbs that interact with the TSHR by measuring their ability to compete with labeled reagents for binding to the TSHR. Therefore these assays do not discriminate between stimulating or blocking activity of the Abs they detect. The second generation assays used labeled TSH to compete with TRAbs in patient sera for binding to a human recombinant TSHR. Instead of TSH, the newer third generation assays substitute a labeled human thyroid-stimulating monoclonal Ab (M22

clone) to compete with patient TRAbs [39]. Naturally, if TRAbs are present, binding of the labeled M22 will be blocked. Therefore modern TRAb assays are often referred to as TBII assays, although it should be noted again that this method does not distinguish between stimulating and blocking autoAb functionalities. These competitive assays are automated and routinely available.

Bioassays have also been used historically for the measurement of thyroid autoAbs, and still are in the case of TSI. In these assays, cell lines have been modified to express a form of the human TSHR. If the receptor is stimulated by the presence of TRAbs in the patient sample, a signaling cascade is initiated within the cell that has a measureable end product (e.g., increased intracellular cAMP). Assay design features, including TSHR makeup and cell lines used, vary by manufacturer or laboratory. Depending on the assay used, the measured response (e.g., cAMP production) is due to stimulation of the TSHR by either TSI alone or the net effect of all TSHR autoAbs. Therefore these TSHR bioassays are a functional measure of TSI activity. By comparing the activity measured in the patient sample with a reference sample (e.g., pooled human serum), the presence of TSI can be expressed as a percentage of basal activity. A high percentage beyond the diagnostic threshold would indicate a positive TSI result. Not surprisingly, these laborious assays are available only in specialty laboratories. Recently, an automated immunoassay specific for TSI has also become available.

References

[1] M. Radu, Physiology of the pituitary, thyroid, parathyroid and adrenal glands, Surgery 32 (2014) 504–512.

[2] C. Fillée, J. Cumps, J.M. Ketelslegers, Comparison of three free T_4 (FT_4) and free T_3 (FT_3) immunoassays in healthy subjects and patients with thyroid diseases and severe non-thyroidal illnesses, Clin. Lab. 58 (2012) 725–736.

[3] A.C. Bianco, D. Salvatore, B. Gereben, M.J. Berry, P.R. Larsen, Biochemistry, cellular and molecular biology, and physiological roles of the iodothyronine selenodeiodinases, Endocr. Rev. 23 (2002) 38–89.

[4] R.P. Peeters, P.J. Wouters, E. Kaptein, H. van Toor, T.J. Visser, G. Van den Berghe, Reduced activation and increased inactivation of thyroid hormone in tissues of critically ill patients, J. Clin. Endocrinol. Metab. 88 (2003) 3202–3211.

[5] R.L. Schmidt, J.S. LoPresti, M.T. McDermott, S.M. Zick, J.A. Straseski, Does reverse triiodothyronine testing have clinical utility? An analysis of practice variation based on order data from a national reference laboratory, Thyroid 28 (2018) 842–848.

[6] G. Hennemann, R. Docter, E.C. Friesema, M. de Jong, E.P. Krenning, T.J. Visser, Plasma membrane transport of thyroid hormones and its role in thyroid hormone metabolism and bioavailability, Endocr. Rev. 22 (4) (2001) 451–476.

[7] J.G. Hollowell, N.W. Staehling, W.D. Flanders, W.H. Hannon, E. W. Gunter, C.A. Spencer, et al., Serum TSH, T(4), and thyroid antibodies in the United States population (1988 to 1994): National Health and Nutrition Examination Survey (NHANES III), J. Clin. Endocrinol. Metab. 87 (2) (2002) 489–499.

[8] C.A. Spencer, M. Takeuchi, M. Kazarosyan, C.C. Wang, R.B. Guttler, P.A. Singer, et al., Serum thyroglobulin autoantibodies: prevalence, influence on serum thyroglobulin measurement, and prognostic significance in patients with differentiated thyroid carcinoma, J. Clin. Endocrinol. Metab. 83 (4) (1998) 1121–1127.

[9] J.R. Garber, R.H. Cobin, H. Gharib, J.V. Hennessey, I. Klein, J.I. Mechanick, et al., Clinical practice guidelines for hypothyroidism in adults: cosponsored by the American Association of Clinical Endocrinologists and the American Thyroid Association, Endocr. Pract. 18 (6) (2012) 988–1028.

[10] D.S. Ross, H.B. Burch, D.S. Cooper, M.C. Greenlee, P. Laurberg, A.L. Maia, et al., American Thyroid Association guidelines for diagnosis and management of hyperthyroidism and other causes of thyrotoxicosis, Thyroid 26 (10) (2016) 1343–1421.

[11] B.R. Haugen, E.K. Alexander, K.C. Bible, G.M. Doherty, S.J. Mandel, Y.E. Nikiforov, et al., 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer, Thyroid 26 (1) (2016) 1–133.

[12] L. Yassa, E.S. Cibas, C.B. Benson, M.C. Frates, P.M. Doubilet, A. A. Gawande, et al., Long-term assessment of a multidisciplinary approach to thyroid nodule diagnostic evaluation, Cancer 111 (2007) 508–516.

[13] P. Caturegli, A. De Remigis, N.R. Rose, Hashimoto thyroiditis: clinical and diagnostic criteria, Autoimmun. Rev. 13 (4–5) (2014) 391–397.

[14] K. Markou, N. Georgopoulos, V. Kyriazopoulou, A.G. Vagenakis, Iodine-induced hypothyroidism, Thyroid 11 (2001) 501–510.

[15] M.V. Rastogi, S.H. LaFranchi, Congenital hypothyroidism, Orphanet J. Rare Dis. 5 (2010) 17.

[16] K. Patil-Sisodia, J.H. Mestman, Graves hyperthyroidism and pregnancy: a clinical update, Endocr. Pract. 16 (1) (2010) 118–129.

[17] J.Y. Tan, K.C. Loh, G.S. Yeo, Y.C. Chee, Transient hyperthyroidism of hyperemesis gravidarum, BJOG 109 (6) (2002) 683–688.

[18] L. De Groot, M. Abalovich, E.K. Alexander, N. Amino, L. Barbour, R.H. Cobin, et al., Management of thyroid dysfunction during pregnancy and postpartum: an Endocrine Society clinical practice guideline, J. Clin. Endocrinol. Metab. 97 (8) (2012) 2543–2565.

[19] J.E. Haddow, G.E. Palomaki, W.C. Allan, J.R. Williams, G.J. Knight, J. Gagnon, et al., Maternal thyroid deficiency during pregnancy and subsequent neuropsychological development of the child, N. Engl. J. Med. 341 (8) (1999) 549–555.

[20] American College of Obstetricians and Gynecologists, Practice Bulletin No. 148: Capitalize Thyroid thyroid disease in pregnancy, Obstet. Gynecol. 125 (4) (2015) 996–1005.

[21] E.K. Alexander, E.N. Pearce, G.A. Brent, R.S. Brown, H. Chen, C. Dosiou, et al., Guidelines of the American Thyroid Association for the diagnosis and management of thyroid disease during pregnancy and the postpartum, Thyroid 27 (3) (2017) 315–389.

[22] L.M. Thienpont, K. Van Uytfanghe, G. Beastall, J.D. Faix, T. Ieiri, W.G. Miller, et al., Report of the IFCC Working Group for Standardization of Thyroid Function Tests; part 2: free thyroxine and free triiodothyronine, Clin. Chem. 56 (6) (2010) 912–920.

[23] A.M. Gronowski, Evaluation of thyroid function during pregnancy: have we taken a wrong turn? Clin. Chem. 64 (3) (2018) 439–441.
[24] https://seer.cancer.gov/statfacts/html/thyro.html. (Accessed November 15, 2018).
[25] F.A. Verburg, M. Luster, C. Cupini, L. Chiovato, L. Duntas, R. Elisei, et al., Implications of thyroglobulin antibody positivity in patients with differentiated thyroid cancer: a clinical position statement, Thyroid 23 (10) (2013) 1211–1225.
[26] S.A. Wells, S.L. Asa, H. Dralle, R. Elisei, D.B. Evans, R.F. Gagel, et al., Revised American Thyroid Association guidelines for the management of medullary thyroid carcinoma, Thyroid 25 (6) (2015) 567–610.
[27] Z. Baloch, P. Carayon, B. Conte-Devolx, L.M. Demers, U. Feldt-Rasmussen, J.F. Henry, et al., Guidelines Committee, National Academy of Clinical Biochemistry, laboratory medicine practice guidelines. Laboratory support for the diagnosis and monitoring of thyroid disease, Thyroid 13 (1) (2003) 3–126.
[28] L.M. Thienpont, K. Van Uytfanghe, S. Van Houcke, B. Das, J.D. Faix, F. MacKenzie, et al., A progress report of the IFCC Committee for Standardization of Thyroid Function Tests, Eur. Thyroid J. 3 (2) (2014) 109–116.
[29] C.A. Spencer, Assay of thyroid hormones and related substances, in: L.J. De Groot, P. Beck-Peccoz, G. Chrousos, et al. (Eds.), Endotext, MDTextcom, Inc., South Dartmouth, MA, 2013. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279113/ (Accessed November 15, 2018).
[30] J.M. Hershman, A.E. Pekary, L. Berg, D.H. Solomon, C.T. Sawin, Serum thyrotropin and thyroid hormone levels in elderly and middle-aged euthyroid persons, J. Am. Geriatr. Soc. 41 (8) (1993) 823–828.
[31] S.S. Tai, L.T. Sniegoski, M.J. Welch, Candidate reference method for total thyroxine in human serum: use of isotope-dilution liquid chromatography-mass spectrometry with electrospray ionization, Clin. Chem. 48 (4) (2002) 637–642.
[32] B. Yue, A.L. Rockwood, T. Sandrock, S.L. La'ulu, M.M. Kushnir, A.W. Meikle, Free thyroid hormones in serum by direct equilibrium dialysis and online solid-phase extraction-liquid chromatography/tandem mass spectrometry, Clin. Chem. 54 (4) (2008) 642–651.
[33] H.E. van Deventer, D.R. Mendu, A.T. Remaley, S.J. Soldin, Inverse log-linear relationship between thyroid-stimulating hormone and free thyroxine measured by direct analog immunoassay and tandem mass spectrometry, Clin. Chem. 57 (1) (2011) 122–127.
[34] S.K. Van Houcke, K. Van Uytfanghe, E. Shimizu, W. Tani, M. Umemoto, L.M. Thienpont, IFCC international conventional reference procedure for the measurement of free thyroxine in serum: International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) Working Group for Standardization of Thyroid Function Tests (WG-STFT), Clin. Chem. Lab. Med. 49 (8) (2011) 1275–1281.
[35] C.A. Spencer, L.M. Bergoglio, M. Kazarosyan, S. Fatemi, J.S. LoPresti, Clinical impact of thyroglobulin (Tg) and Tg autoantibody method differences on the management of patients with differentiated thyroid carcinomas, J. Clin. Endocrinol. Metab. 90 (10) (2005) 5566–5575.
[36] M.M. Kushnir, A.L. Rockwood, W.L. Roberts, D. Abraham, A.N. Hoofnagle, A.W. Meikle, Measurement of thyroglobulin by liquid chromatography-tandem mass spectrometry in serum and plasma in the presence of antithyroglobulin autoantibodies, Clin. Chem. 59 (6) (2013) 982–990.
[37] A.N. Hoofnagle, J.O. Becker, M.H. Wener, J.W. Heinecke, Quantification of thyroglobulin, a low-abundance serum protein, by immunoaffinity peptide enrichment and tandem mass spectrometry, Clin. Chem. 54 (11) (2008) 1796–1804.
[38] J.L. Powers, F.G. Strathmann, J.A. Straseski, Thyroglobulin antibody screen prior to mass spectrometry provides measurable cost savings and optimal laboratory utilization, Am. J. Clin. Pathol. 14 (3) (2017) 309–314.
[39] B.R. Smith, J. Bolton, S. Young, A. Collyer, A. Weeden, J. Bradbury, et al., A new assay for thyrotropin receptor autoantibodies, Thyroid 14 (10) (2004) 830–835.

Chapter 41

Disorders of the adrenal cortex and medulla

David N. Alter

Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Explain the regulation and actions of aldosterone, cortisol, adrenal androgens, and catecholamines.
- Diagnose conditions involving excesses of aldosterone, cortisol, adrenal androgens, or catecholamines.
- Diagnose conditions involving deficiencies of aldosterone and/or cortisol.

Introduction

The adrenal glands influence several different organ systems. They are located suprarenally, weigh about 4–6 g each, and are structurally and functionally split into cortical (cortex/outer) and medullary (inner) layers. Embryologically, the adrenal cortex arises from the mesoderm and shares biochemical pathways associated with the urogenital system. Similarly, the adrenal medulla is of neuroectodermal (e.g., neural crest) origin and shares biochemical pathways similar to the nervous system. Products of the adrenal cortex consist of mineralocorticoids (aldosterone), glucocorticoids (cortisol), and adrenal androgens [dehydroepiandrosterone (DHEA) and androstenedione]. The adrenal medulla is responsible for catecholamine (e.g., epinephrine) production. Adrenal cortex and medullary hormones and their regulation are listed in Table 41.1.

Similar to other endocrine gland dysfunction, disease can present as adrenal hypofunction, hyperfunction, or a combination of both [e.g., congenital adrenal hyperplasia (CAH)]. Endocrine tumors may present with or without functional changes.

In most situations, biochemical evaluation for endocrine disorders should not be performed unless the history and clinical presentation are compatible with the suspected disorder.

Adrenal cortex

The adrenal cortex is composed of three layers; each responsible for production of specific physiologic steroids, respectively (from outer to inner): zona glomerulosa [mineralocorticoids (aldosterone)], zona fasciculata [glucocorticoids (cortisol) with small amounts of androgens], and zona reticularis (androgens DHEA and androstenedione; see Fig. 41.1 and Table 41.2).

TABLE 41.1 Adrenal gland.

Adrenal layer	Primary hormone(s)	Regulation
Cortex		
Zona glomerulosa	Aldosterone Deoxycorticosterone	Renin–angiotensin–aldosterone system
Zona fasciculata	Cortisol	Hypothalamic–pituitary–adrenal axis
Zona reticularis	Dehydroepiandrosterone Androstenedione	
Medulla	Epinephrine	Sympathomimetic nervous system

HPA, Hypothalamic–pituitary–adrenal.

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00041-7

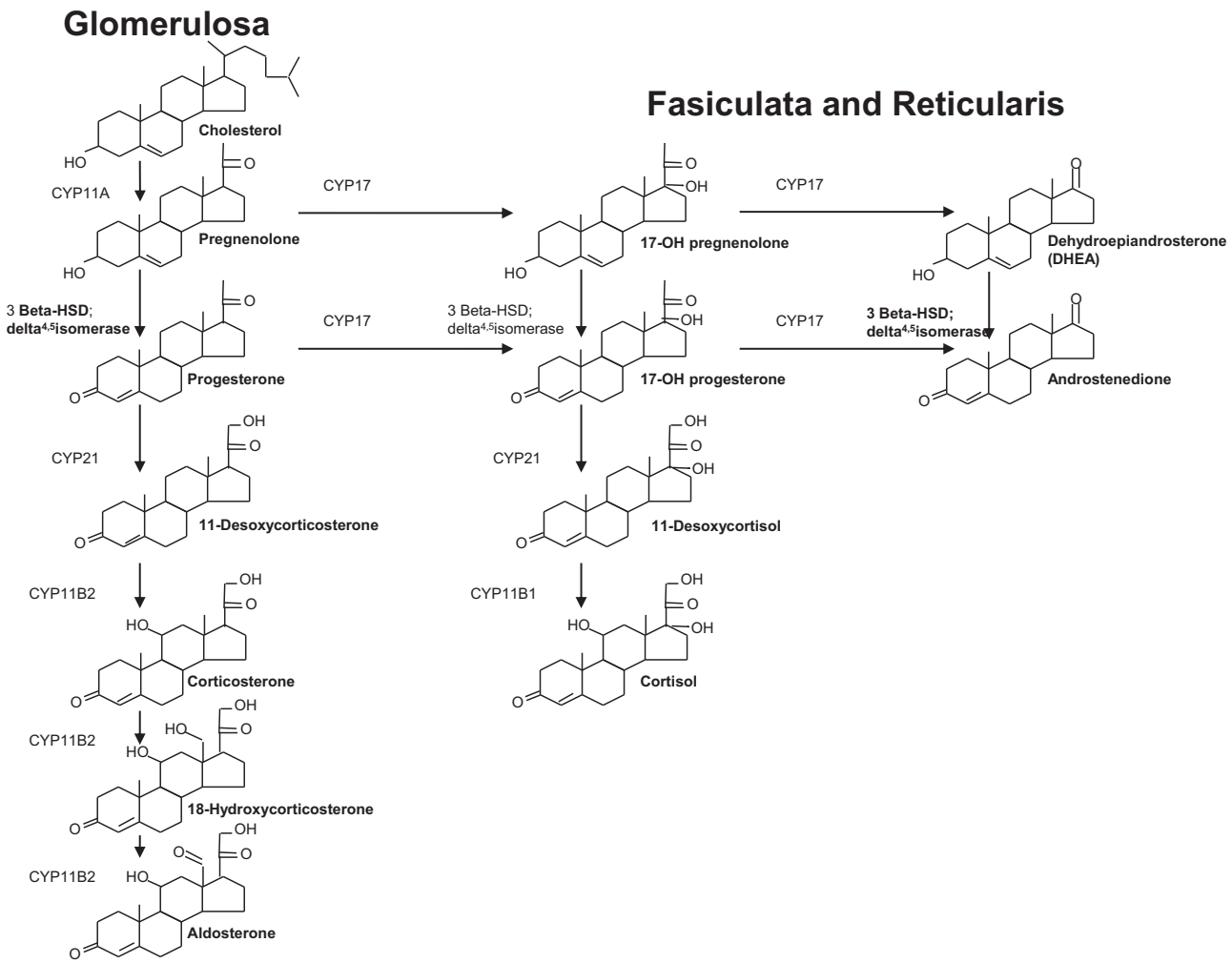


FIGURE 41.1 Adrenal hormone synthetic pathways.

Mineralocorticoids

Aldosterone is the primary mineralocorticoid and its secretion/inhibition is associated with sodium concentrations as well as blood pressure/renal perfusion. Aldosterone increases sodium concentrations via the stimulation of sodium reabsorption channels with its secretion controlled through the renin−angiotensin−aldosterone system (RAAS): an interplay of baroreceptors located in the glomerular afferent arterioles working in conjunction with sodium chemoreceptors located in the macula densa of the juxtaglomerular (JG) apparatus. The RAAS is a dynamic process that mediates renin secretion from the JG cells dependent on volume and sodium concentration. In low volume/sodium (physiologic and pathophysiologic) states, basal renin secretion increases leading to increased renin cleavage (as well as all the subsequent steps) of angiotensinogen (485 aa) into angiotensin I (10 aa), which is then converted into angiotensin II (8 aa) via angiotensin-converting enzyme (ACE; located in the lungs). Finally, increased production of Angiotensin II, in turn, stimulates an increase in aldosterone synthesis and release. Aldosterone stimulates sodium reabsorption at the proximal convoluted tubule and vasoconstriction along with increased thirst, increased antidiuretic hormone (ADH) release, and catecholamine release (Fig. 41.2). Ultimately, as described, the RAAS promotes water retention and increases blood volume, with subsequent maintenance of blood pressure. Unsurprisingly, aldosterone excess disorders result in hypertension. In addition, deoxycorticosterone (DOC), an intermediate in the aldosterone synthesis pathway, also possesses systemic mineralocorticoid activity, and in scenarios to be described, DOC excesses will also be presented as hypertension.

Disorders of mineralocorticoid excess

Refractory and/or unexplained hypertension is the underlying clinical feature associated with mineralocorticoid

TABLE 41.2 Mineralocorticoid excess disorders.

Increased aldosterone	*Primary (hyporeninemic)*
	Bilateral IAH (60%–65%)
	Aldosterone-producing adrenal adenoma[APA "aldosteronoma" unilateral (30%–35%)];
	Adrenal hyperplasia (2%)
	Aldosterone producing adrenocortical carcinoma (<1%)
	Familial hyperaldosteronism
	FH Type I GRH <1%
	FH Type II APA or IAH linked to chromosome 7p12 <6%
	FH Type III KCNJ5 massive adrenal hyperplasia of childhood <1%
	Secondary (hyperreninemic)
	Renin-secreting tumors
	Renovascular hypertension
	Compensatory (nephrosis, cirrhosis, and congestive heart failure)
	Renal tubular disorders (e.g., Bartter and Gitelman syndrome)
Normal/low aldosterone	*Hyperdeoxycorticosterone*
	DOC-secreting adrenal adenoma
	11 beta-hydroxylase (CYP11B1) deficiency
	17-hydroxylase (CYP17) deficiency
Glucocorticoid excess	Cushing syndrome
	Cortisol resistance
End-organ defects	Apparent mineralocorticoid excess: HSD11B2 deficiency or glycyrrhizic acid-induced inhibition
	Liddle syndrome (type 1 pseudohyperaldosteronism)

APA, Aldosterone producing adenoma; *DOC*, deoxycorticosterone; *GRH*, glucocorticoid remediable hyperaldosteronism; *HSD11B2*, hydroxysteroid dehydrogenase-11-beta; *IAH*, idiopathic adrenal hyperplasia.

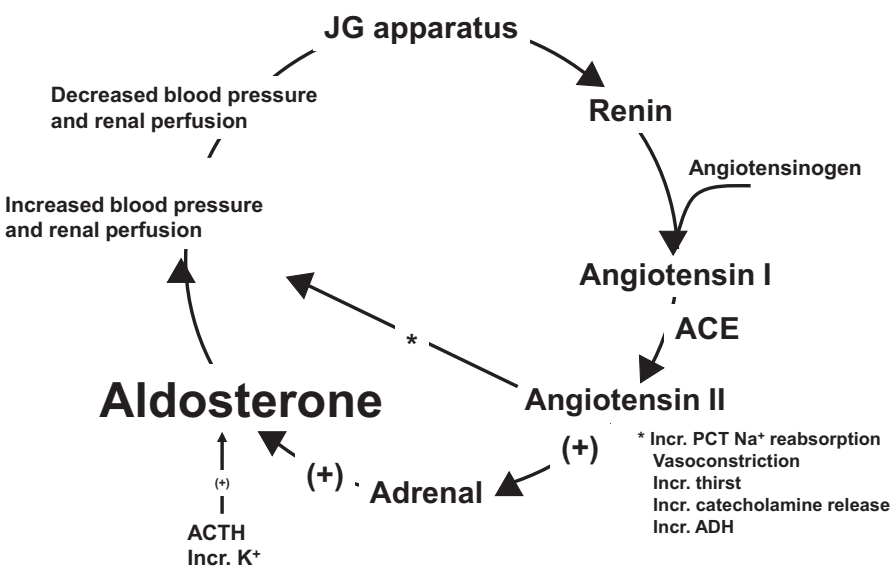


FIGURE 41.2 Aldosterone regulation.

excess or related disorders (Table 41.2). Most are due to aldosterone elevations; however, excess DOC [secondary to 17-alpha hydroxylase and 11-beta hydroxylase (11B-OH) deficiencies] may also be responsible as well as those referred to as apparent mineralocorticoid excess (AME) disorders.

AME disorders are associated with increased mineralocorticoid receptor (MR) activity in the absence of excess mineralocorticoid production. For example, in individuals with 11-beta-hydroxysteroid dehydrogenase (11B-HSD2) deficiency, cortisol is not converted into cortisone with the increased cortisol concentrations stimulating the MR, resulting in aldosterone-like activity. 11B-HSD2 deficiency presents similar to hyperaldosteronism [hypertension, hypokalemia, metabolic alkalosis, and low plasma renin activity (PRA)] but does not have the associated mineralocorticoid elevations. A very rare cause of AME is Liddle's syndrome, an autosomal dominant upregulation of the epithelial sodium channel (ENaC). In this condition, sodium reabsorption occurs regardless of a low/blunted aldosterone response with patient presentation significant for young age of onset, excess sodium retention, and low/normal aldosterone levels. An AME-like disorder involves excess consumption of black licorice containing glycyrrhetinic acid. Lastly, pathologic cortisol elevations from other etiologies not AME (e.g., excess exogenous glucocorticoid intake, endogenous cortisol excess, and cortisol resistance) can also trigger excessive mineralocorticoid activity via activation of the MR.

Disorders of mineralocorticoid excess: laboratory testing

Aldosterone can be measured in plasma or in urine following a timed urine collection. Renin is measured as PRA or plasma renin concentration (PRC). Both assays measure PRA based on renin's ability to convert angiotensinogen into angiotensin I; however, PRA assays utilize endogenous (in vivo) angiotensinogen and PRC assays utilize exogenous (in vitro) angiotensinogen (provided with the assay). Consequently, PRC assays are subject to less variability of result, because a standardized angiotensinogen concentration is used. For the rest of this discussion, "PRA" will be used to represent both assay types. PRA has numerous preanalytic variables such as posture (increased with standing), time of day (diurnal variation), age (PRA decreases with age), dietary sodium (PRA decreases with sodium intake), and other medications (renin is increased if on a patient is on a diuretic). Due to the influence of extra laboratory variables on renin measurements, care must be taken with result reporting and interpretation. An aldosterone/PRA ratio $>20-25$ is considered consistent with primary hyperaldosteronism

Confirmation studies for primary hyperaldosteronism include the failure of aldosterone to be suppressed by one of the following volume expanders: fludrocortisone, oral sodium, or intravenous saline administration. Volume expansion normally suppresses aldosterone levels; therefore failure to suppress aldosterone indicates autonomous aldosterone production. Great care must be taken during saline infusion or oral salt loading as blood pressure may rise and (if preexisting moderate-to-severe hypertension present) can precipitate a hypertensive crisis or the development of pulmonary edema if there is a preexisting marginal cardiac function.

There is also the captopril suppression test, as captopril inhibits ACE; therefore captopril administration should inhibit aldosterone production. Failure to inhibit plasma aldosterone levels using this approach is also consistent with primary aldosteronism, but this test does not appear to carry the same risks as the previously mentioned suppression tests.

When primary hyperaldosteronism is confirmed, CT or MRI in conjunction with R/L adrenal and R/L peripheral vein samplings of aldosterone and cortisol is performed to distinguish bilateral [idiopathic adrenal hyperplasia (IAH)] from unilateral [aldosterone producing adenoma (APA)] disease. Right and left adrenal vein measurements of cortisol and aldosterone are used to confirm adrenal vein cannulation and to determine laterality (if any) of the aldosteronism. Aldosterone/cortisol ratios are calculated to correct for the dilutional effect of blood from the inferior phrenic vein flowing into the left adrenal vein. These measurements are interpreted thusly: unilateral aldosteronism (APA) if the aldosterone/cortisol ratio from side to side is $>4:1$; bilateral (IAH) if the ratio is $\leq 3:1$; and indeterminate if the results are between 3.1 and 3.9 [1–3].

Glucocorticoid remediable hyperaldosteronism (GRH) is suggested by early-onset hyperaldosteronism with a family history of similarly affected individuals in an autosomal dominant pattern. It occurs due to a crossing-over defect between the genes encoding aldosterone synthase and 11B-OH. Both genes are located near each on chromosome 8 such that they can fuse together forming an aldosterone synthase sensitive to adrenocorticotropic hormone (ACTH). The "chimeric" gene converts cortisol into 18-hydroxy and 18-oxometabolites with increased urinary tetrahydro-18-oxocortisol. Therefore in patients with GRH, the ratio of tetrahydro-18-oxocortisol to tetrahydroaldosterone is increased. In practice, reduction in blood pressure secondary to dexamethasone administration supports the diagnosis of GRH. AME is diagnosed by increased ratios of cortisol to cortisone and increased urinary tetrahydrocortisol to tetrahydrocortisone ratios [4].

Disorders of mineralocorticoid deficiency

Hypoaldosteronism biochemically presents as hyponatremia and hyperkalemia with hydrogen ion-retention causing acidosis. Clinically, salt-craving is associated with mineralocorticoid deficiency as a protective response. Aldosterone deficiency is commonly seen in the cases of primary adrenal insufficiency associated with a coexistent cortisol deficiency. Rarely, primary isolated aldosterone deficiency results from loss-of-function mutations in aldosterone synthase (CYP11B2). Renin deficiency is a rare condition, but can be seen in people with long-standing diabetic nephropathy. The syndrome of hyporeninemic hypoaldosteronism (SHH, type 4 distal renal tubular acidosis) occurs in late middle age, and is associated with diabetes and renal insufficiency secondary to tubulointerstitial disease. These patients are hyperkalemic with a hyperchloremic metabolic acidosis. Loss-of-function mutations in the MR cause pseudohypoaldosteronism, where aldosterone is present but there is resistance to its actions [5].

Disorders of mineralocorticoid deficiency: laboratory testing

Hyponatremia, hyperkalemia, acidosis, hypovolemia, hypotension, and possibly shock are the diagnostics for mineralocorticoid deficiency in the setting of suspected primary adrenal insufficiency [Addison disease, autoimmune polyglandular syndrome (APS), Waterhouse–Friderichsen syndrome, etc.]. Aldosterone can be measured in plasma in search of deficiency; however, the clinical findings in the appropriate clinical scenario are usually sufficient to make the diagnosis of mineralocorticoid insufficiency. An elevated PRA and a low aldosterone concentration support the diagnosis of mineralocorticoid deficiency. SHH is notable for its clinical presentation as well as low renin and aldosterone levels that are not responsive to stimulatory testing. Failure of aldosterone to rise in response to standing or furosemide indicates aldosterone insufficiency. Great care must be taken, because acute fluid loss from a furosemide-induced diuresis can precipitate acute hypovolemia and an Addisonian crisis. Although not a test of mineralocorticoid function, the finding of adrenal autoantibodies [either adrenal cytoplasmic autoantibodies detected by indirect immunofluorescence or 21-hydroxylase (21-OH) autoantibodies] identifies that a patient has autoimmune adrenalitis and will be at high risk of adrenal insufficiency due to Addison disease.

Glucocorticoids

Cortisol is the primary glucocorticoid secreted from the adrenal cortex and has numerous actions throughout multiple organs and organ systems with its synthesis and secretion regulated centrally by the hypothalamic–pituitary–adrenocortical–cortisolll axis. Glucocorticoid actions include elevated blood glucose, increased cardiac inotropism/vascular tone, and impaired monocyte/macrophage/CD4-T cell activity. Cortisol has a diurnal release with concentrations highest in the early morning and lowest near midnight; it is released in response to both physiologic and pathologic stimuli. Physiologically, stress situations such as infection, sleep deprivation, and caloric restriction can also trigger release. In contrast, nonpathologic/nonphysiologic stimuli can also affect cortisol release, for example, music (decrease), dancing (decrease), and caffeine intake (increase). Pathologic changes in cortisol release such as Cushing syndrome and adrenal insufficiency will be discussed below.

Under the regulation of neurotransmitters, the hypothalamus produces corticotropin-releasing hormone (CRH; 41 aa) delivered to the anterior pituitary via the hypothalamic–pituitary portal system. CRH binds to anterior pituitary corticotroph receptors, triggering the synthesis of proopiomelanocortin, which upon release is cleaved into ACTH (39 aa; corticotropin) as well as additional fragments that include alpha, beta, and gamma melanocyte stimulating hormones, beta-lipotropin, and beta-endorphins. ACTH release can also be caused by high concentrations of ADH and inflammatory cytokines such as IL-1, IL-6, and tumor necrosis factor-alpha. ACTH binds to receptors on cells of the zona fasciculata and reticularis resulting in increased cortisol production and release.

TABLE 41.3 Cushing syndrome.

Weakness (steroid myopathy especially affecting the proximal muscles)
Back pain (osteoporosis, fractures, and compressed vertebrae)
Bruisability (catabolism of connective tissue)
Psychiatric problems (emotional lability, depression, or psychosis)
Weight gain
Centripetal obesity
Moon facies
Buffalo hump
Supraclavicular fat pad
Plethora
Hypertrichosis
Hirsutism
Acne
Increased susceptibility to infection and hypertension

Disorders of glucocorticoid excess

Cushing syndrome refers to the signs, symptoms, and sequelae (Table 41.3) associated with hypercortisol states, whereas Cushing disease specifically indicates that the source of elevated cortisol is due to an ACTH-producing pituitary adenoma. Cushing syndrome is typically sorted into ACTH-dependent and ACTH-independent causes. ACTH-dependent causes are due to excess ACTH, whereas, ACTH-independent causes are not.

ACTH-dependent (Cushing disease, ectopic ACTH, and ectopic CRH):

- Cushing disease (68%) may occur as a consequence of an anterior pituitary corticotroph adenoma (corticotropinoma). Hyperpigmentation may be present when ACTH levels exceed 300 pg/mL. Additional possible sequelae include hypokalemic alkalotic hypertension, and low eosinophil count and varying degrees of elevated blood glucose concentrations.
- Ectopic ACTH syndrome (6%−15%) is typically seen as a component of certain tumor-related paraneoplastic syndromes, including small-cell lung cancers.
- Ectopic CRH syndrome is very rare and has been associated as a paraneoplastic syndrome of numerous malignancies (bronchial, thymic neuroendocrine, small-cell lung, medullary thyroid, pheochromocytoma, gangliocytoma, prostate carcinoma, and ganglioneuroblastoma).

ACTH-independent [adrenal Cushing syndrome, adrenal carcinoma, pseudo-Cushing, primary pigmented nodular adrenal disease (PPNAD), carney syndrome, McCune-Albright syndrome, and exogenous glucocorticoid administration]:

- Adrenal Cushing's syndrome (8%−19%) occurs when a cortisol-secreting adrenocortical adenoma is present where a single adrenal gland weighs between 25 and 40 g.
- Adrenal hyperplasia (2%) can manifest as bilateral adrenal hyperplasia, with maximum adrenal weights between 30 and 50 g. Bilateral adrenal hyperplasia can be micronodular (<3 mm) or macronodular (>3 mm). Macronodular adrenocortical disease is seen in older adults, whereas micronodular (PPNAD) typically presents as an inherited condition in children, for example, Carney complex. "Pigmented" in that the adrenal gland, in these cases, is studded with dark brown and black nodules.
- Adrenal carcinoma (6%−7%) is similar to its adenomatous cousin except that they typically weigh more than 100 g. Unfortunately, when discovered, they are already metastatic.
- "Pseudo-Cushing" syndrome occurs secondary to elevated cortisol levels from drugs, depression, or underlying illness.
- Exogenous glucocorticoid administration may very well account for most cases of Cushing syndrome because of its endemic use in a variety of medical conditions.

Disorders of glucocorticoid excess: laboratory testing

The diagnosis of Cushing syndrome should proceed stepwise from the clinical suspicion to confirmation of a hypercortisol state to the determination of etiology. Cortisol excess is identified using one of three screening tests: 24-hour urine-free cortisol (UFC) concentrations, the overnight low-dose dexamethasone suppression test (LDDST), or late night salivary cortisol measurements. Due to cortisol's diurnality, random levels are of little clinical value; however, a midnight value cortisol of <2 μg/dL essentially rules out Cushing syndrome with sensitivities in the low 90s and specificities approaching 100%. Some investigators would use a less restrictive cutoff of <5 μg/dL to rule out Cushing syndrome.

UFC (24-hour urine specimen) is an integrated measure of cortisol production and is usually two to three times above the upper limit of the reference interval in Cushing syndrome. Preanalytically, errors in collection (i.e., significantly longer or shorter than 24 hours) can erroneously alter the result. Simultaneous urine creatinine measurements are recommended to ensure completeness of collection with the expected concentrations as denoted (men: 14−26 mg per kg of body mass/24 hour; women: 11−20 mg per kg of body mass/24 hour). Elevated UFC levels have a 95%−100% sensitivity for the diagnosis of Cushing syndrome.

Salivary cortisol levels, like serum cortisol, reach a nadir in the late evening; however, unlike serum specimens, a salivary specimen can be collected at home and does not require refrigeration. Preanalytically, salivary specimens have their own issues. According to one reference laboratory, the sample should not be collected within 60 minutes of eating, no sooner than 12 hours after alcohol intake, and should not be collected immediately after brushing or flossing. An appropriate kit to collect saliva is required (e.g., patients should not simply spit and seal into a cup). Preanalytically, salivary cortisol measurements increase with age, diabetes, and blood pressure. Finally, LC-MS is recommended as the analytic modality, as cortisol immunoassays are subject to false-positive results due to cortisone (converted from cortisol into the salivary glands) cross reactivity. Diagnostically, data suggest that salivary cortisol testing can detect mild hypercortisolism missed by UFC and has reported sensitivities and specificities in the mid to high 90 seconds.

In the overnight LDDST, 1 mg of dexamethasone is taken orally between 2200 and midnight followed by a cortisol measurement the next morning between 0800 and 0900. A normal response would be less than 1.8 μg/dL confirming ACTH suppression, with elevated values indicating nonsuppression, that is, pathologic hypercortisolism. Some investigators use a less restrictive cutoff of <5 μg/dL to rule out Cushing syndrome. Elevated cortisol values (indicating nonsuppression) have a sensitivity approaching 100% for the diagnosis of Cushing syndrome. False-positive and false-negative LDDST results can occur to drug metabolism defects (induction/inhibition of CYP 3A4 system). False-positive results can also be seen in elevated cortisol-binding globulin states (e.g., pregnancy, estrogens, and tamoxifen), and liver disease can result in false-negative results. There is also a 48-hour version of the LDDST. In that test, there is a similar sensitivity, but its specificity approaches 100% (overnight LDDST; specificity approximately 88%). There is also a high-dose variation of the LDDST where 8 mg is administered. The extended LDDST and high-dose variations are used in cases where there are equivocal or unexpected LDDST results. The etiology of confirmed hypercortisolism can be defined using the low-dose–high-dose dexamethasone suppression test, which is described in the chapter discussing pituitary disorders (Table 41.4).

For patients with equivocal physical findings and none/mild biochemical changes, a combined dexamethasone–cortisol releasing hormone test has been proposed. Essentially, dexamethasone is administered every 6 hours for 48 hours followed by one CRH dose at 46 hours with cortisol then measured 5 minutes before, at administration, and 15 minutes afterward. Sensitivity and specificity in the high 90 seconds were reported. Its use has been suggested as a way to sort out pseudo-Cushing and Cushing syndromes. The insulin tolerance test has also been used to sort out Cushing syndrome from pseudo-Cushing states (Chapter 33: Evaluation of exocrine pancreatic function) for more information on insulin-tolerance testing.

In general, when evaluating possible hypercortisolism; if the initial screening test indicates hypercortisolism, confirmatory testing should be performed. If the second test is normal, a third screening test should be performed. If two out of three of the screening tests are abnormal, hypercortisolism is confirmed. The challenge then is to determine the etiology of the hypercortisolism.

After hypercortisolism is confirmed, ACTH testing and radiologic studies should be performed to determine its etiology and location (if necessary). Elevated versus depressed ACTH levels distinguish ACTH-dependent and ACTH-independent etiologies with equivocal results requiring a CRH stimulation test to assess pituitary response. Radiologic studies are key in detecting adrenal tumors or hyperplasia; however, hyperplasia can be seen in ACTH-dependent and ACTH-independent causes.

Bilateral inferior petrosal sinus with peripheral vein sampling (a unique combination of radiology and clinical chemistry) for ACTH with/without the CRH stimulation test can determine if the elevated ACTH is of pituitary or peripheral origin. If the elevation in ACTH concentrations is attributable to the pituitary gland, the ratio of pituitary to peripheral ACTH level is >3:1. In contrast, if the source of elevated ACTH is adrenal or ectopic, the ratio of pituitary to peripheral ACTH levels is <3:1. This test is highly sophisticated and requires interventional radiologists with the appropriate skill and experience to perform the test, because erroneous results will occur if the wrong region is cannulated. Additional complications associated with this sampling include increased blood loss, risk of infection, and sinus perforation. Simultaneous prolactin measurement of these specimens is recommended to confirm pituitary and nonpituitary localization.

As discussed, CRH, released by the hypothalamus, triggers pituitary ACTH release, which stimulates adrenal gland production of cortisol. The CRH stimulation test in combination with the LDDST can be used to sort out pseudo-Cushing from Cushing syndrome. In patients with pseudo-Cushing, serum cortisol would suppress to <1.4 μg/dL, where it would be higher in patients with Cushing syndrome. The CRH test, by itself, can also be used to sort out central versus ectopic ACTH etiologies. In the cases of Cushing disease, there will be elevations in plasma ACTH and cortisol, whereas, in the cases of ectopic ACTH, elevations in ACTH and cortisol levels will not occur. CRH used in conjunction with arginine, vasopressin, and desmopressin will exaggerate the ACTH response allowing for differentiation of central versus ectopic causes of ACTH [6–8].

Disorders of glucocorticoid deficiency (adrenal insufficiency)

Adrenal insufficiency [9] is due to adrenal, central, and exogenous causes. It is also characterized as primary (Addison disease), secondary, and, in some texts, tertiary. For this discussion, primary refers to disorders specific to the adrenal gland, secondary typically refers to pituitary (and in some sources) hypothalamic causes, and tertiary, when used, refers to adrenal insufficiency secondary to suppression from exogenous glucocorticoid use (as well as hypothalamic in other sources). Regardless of cause, adrenal insufficiency clinically presents as weakness, malaise, nausea, vomiting, diarrhea, dehydration, weight loss, hypoglycemia, hyponatremia, hyperkalemia, and acidosis resulting from a combination of coexistent aldosterone and cortisol deficiencies. Acute adrenal insufficiency

TABLE 41.4 Dexamethasone suppression test.

			Abnormal response (failure of suppression)	Sens	Spec
Low-dose DST					
Overnight, 1-mg test	Most common screen; DX of endogenous CS	Dexamethasone 1-mg orally 11 p.m.–12 a.m. Cortisol levels between 8and 9 a.m.	Cortisol 1.8 μg/dL (>50 nmol/L)	95%	86%
Two-day, 2-mg test		Dexamethasone 0.5 mg is administered orally every 6 h (9 a.m., 3 p.m., 9 p.m., and 3 a.m.) for two days (total dose 4 mg). Cortisol level 6 hours after last dose (9 a.m.)	Cortisol 1.8 μg/dL (>50 nmol/L)	97%–100%	>99%
High-dose DST					
Overnight, 8-mg test	Ectopic versus pituitary Overproduction of ACTH in Cushing disease (but not ectopic tumors) can undergo partial or full suppressed by high doses of dexamethasone	Overnight, 8-mg test Baseline a.m. morning cortisol Oral dexamethasone 8 mg between 11 p.m. and midnight Repeat serum cortisol is drawn the next morning (between 8 and 9 a.m.)		60%–100% if 50% suppression	60%–100% if 50% suppression
Two-day, 8-mg test		Day1 Baseline a.m. morning serum cortisol or 24-h urine-free cortisol is obtained Day 2 and 3, oral dexamethasone 2 mg every 6 h (9 a.m., 3 p.m., 9 p.m., and 3 a.m.) is given (total eight doses; 16 mg) with simultaneous collection of a urine sample for urine-free cortisol. Serum cortisol levels are checked 6 h after the last dose (9 a.m.)			
Dexamethasone-CRH test	Glucocorticoid suppression of the HPA axis can be overcome by CRH stimulation in Cushing disease and not in pseudo-Cushing syndrome	Dexamethasone 0.5 mg every 6 h (12 p.m., 6 p.m., 12 a.m., and 6 a.m.) is given orally for 48 h. Two hours after the last dose of dexamethasone, intravenous CRH 1 mcg/kg is administered (8 a.m.). Serum cortisol is drawn 15 min later		Cortisol > 1.4 μg/dL at 15 min CD 90–100 Cortisol > 3.8 μg/dL at 15 min CD 94	Cortisol > 1.4 μg/dL at 15 min CD 50–100 Cortisol > 3.8 μg/dL at 15 min CD 100

ACTH, Adrenocorticotropic hormone; *CRH*, corticotropin releasing hormone; *CS*, cushing syndrome; *DST*, dexamethasone suppression test; *DX*, diagnosis; *HPA*, hypothalamic–pituitary–adrenal.

(major trauma, surgery, gastroenteritis, and hemorrhage) will present with severe symptoms such as obtundation, hypotension, and shock (Addisonian crisis). Without rapid and appropriate fluid restoration and glucocorticoid replacement, an Addisonian crisis can be fatal.

Historically, and in developing countries, adrenal insufficiency was/is due to infection (first reported secondary to tuberculosis). In developed countries, the most common cause of Addison disease is autoimmune destruction (adrenalitis). It can occur as an isolated condition or as part of an APS. In 90% of recent onset patients, autoantibodies to the 21-OH enzyme were identified and their presence carries a 48% risk of Addison's disease. The remainder of the reported cases of Addison disease is attributed to APS.

APS type I is defined by the presence of at least two out of three of the following findings: hypoparathyroidism, mucocutaneous candidiasis, and Addison disease or adrenal autoantibodies. Associated findings include primary ovarian failure, alopecia, vitiligo, ectodermal dysplasia (e.g., dental enamel hypoplasia and nail dystrophy), red cell aplasia, IgA deficiency, malabsorption, autoimmune hepatitis, progressive myopathy, and type 1 diabetes (T1DM). APS type I is also known as the autoimmune polyendocrinopathy, candidiasis, ectodermal dystrophy syndrome. Onset is usually in childhood with APS type I inherited as an autosomal recessive trait (not gender specific). At the molecular level, it is due to a loss-of-function mutation in the autoimmune regulator (*AIRE*) gene I. AIRE has a role in the expression of self-antigens in the thymus necessary for the development of tolerance to self.

APS type II is identified by the presence of Addison disease or adrenal autoantibodies and either autoimmune thyroid disease (AITD), T1DM, or both T1DM and AITD. In contrast to APS type I, type II disease is inherited in a polygenic fashion and is more common in women than in men with onset later than APS type I. Associated findings include primary ovarian failure, pernicious anemia, hypophysitis, autoimmune hepatitis, vitiligo, alopecia, dermatitis herpetiformis, IgA deficiency, celiac disease, immune thrombocytopenic purpura, myasthenia gravis, stiff person syndrome, and Parkinson disease.

Following autoimmunity, the next most common cause of Addison disease is infections such as tuberculosis, HIV, cytomegalovirus, herpes simplex, syphilis, fungal infections [e.g., blastomycosis (North and South American), *Histoplasma capsulatum*, cryptococcosis, sporotrichosis, and coccidiomycosis], or sarcoidosis. Other causes of Addison disease include adrenoleukodystrophy, CAH, Wolman disease, Smith–Lemli–Opitz syndrome, and adrenal infiltrative diseases, for example, amyloidosis, hemochromatosis, or cancer.

Congenital adrenal hyperplasia

CAH is a number of autosomal recessive disorders characterized by defective genes coding for numerous enzymes in the adrenal corticosteroid biosynthetic pathway. Fig. 41.3 details all of the inborn errors associated with adrenal steroid biosynthesis with those associated with CAH denoted.

21-OH (CYP21A2) deficiency is the most common type of CAH. This enzymatic deficiency occurs in 1 out of 15,000 births and represents up to 95% of all CAH cases. 21-OH is responsible for the conversion of 17-hydroxyprogesterone (17-OHP) into 11-deoxycortisol. Enzyme deficiency will result in elevated levels of 17-OHP and decreases in 11-deoxycortisol concentrations; 75% patients clinically present with mineralocorticoid deficiency. The disorder is categorized into three categories: classic salt wasting, simple virilizing, and nonclassic.

Classic salt-wasting 21-OH deficiency presents at birth or soon after, and can be fatal depending on degree of salt wasting. Females with the disorder display genital ambiguity (i.e., virilization of the external genitalia). Individuals deficient with 21-OH clinically present with hyponatremia, hyperkalemia, and failure to thrive.

Simple virilizing 21-OH deficiency causes virilization of the female fetus and may not be noticed initially in males. Males can show early (2–4 years) virilization and precocious puberty. Nonclassic 21-OH deficiency presents later in life with females showing virilizing features [10,11]. Diagnosis for both classic and nonclassic types depends on clinical presentation (genital ambiguity, salt wasting, and hypotension) and can be confirmed with elevated 17-OHP levels. However, classic 21-OH deficiency will have markedly elevated 17-OHP (>1200 ng/dL with most neonates >3500 ng/dL) levels; whereas, nonclassic not so much. Nonclassic 21-OH deficiency in children is associated with elevated early morning (>200 ng/dL) 17-OHP levels ($>95\%$ sensitivity and specificity). For adults, a 17-OHP value >200 ng/dL strongly suggests the disorder with an exaggerated response (>1500 ng/dL) post-ACTH stimulation considered confirmatory.

Prior to the development of newborn screening programs to detect CAH, more girls than boys were diagnosed with CAH. With current screening programs that measure 17-OHP in dried blood spots, the number of boys and girls diagnosed with CAH is equivalent. 11B-OH deficiency accounts for about 7% of cases of CAH with features similar to the simple virilizing form of 21-OH deficiency. Clinically, it has been associated with excess DOC activity showing hypernatremia, hypokalemia, and hypertension. In addition, rare etiologies of CAH are discussed in the section on androgen disorders.

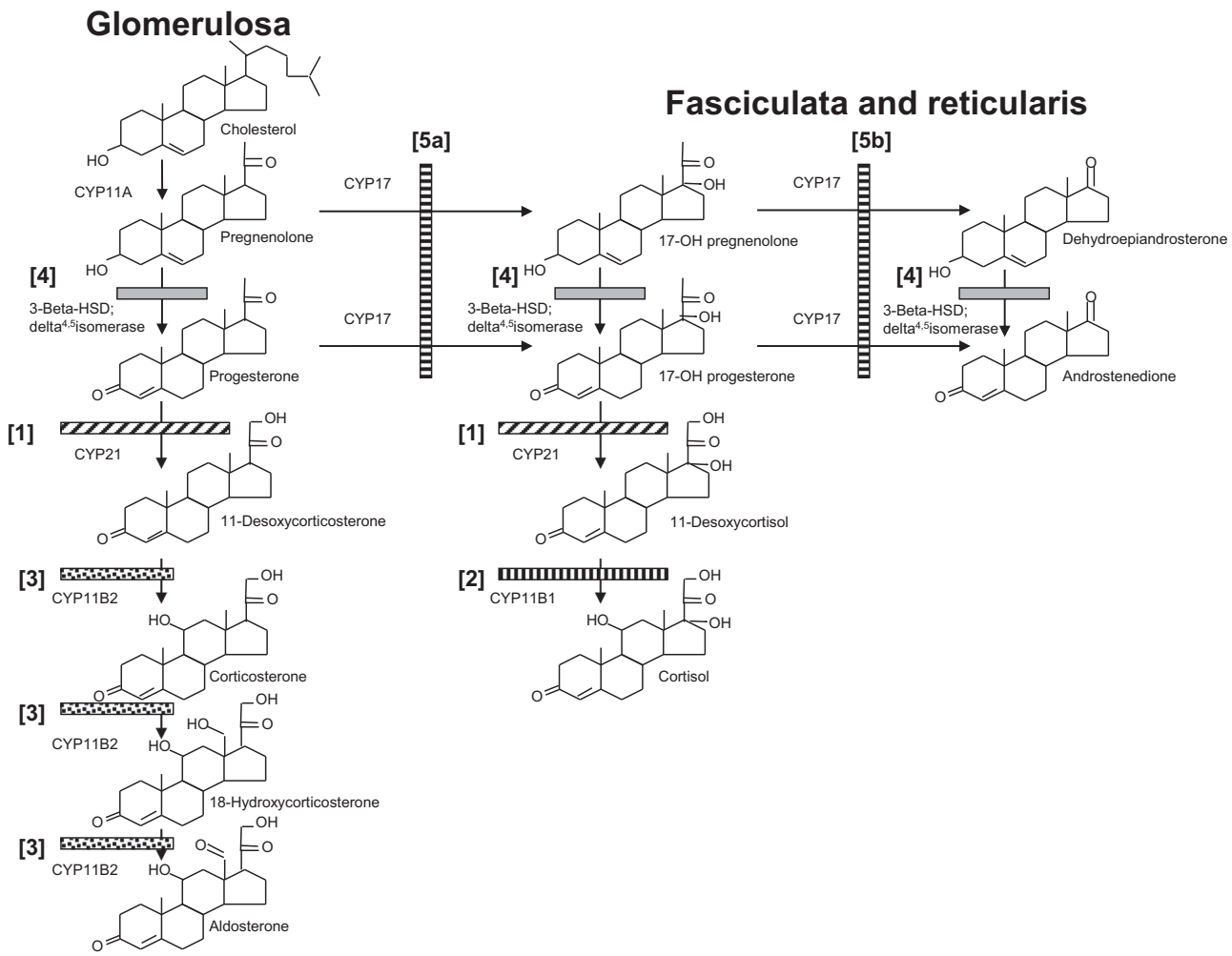


FIGURE 41.3 Inborn errors in adrenal steroid biosynthesis. [1] 21 alpha-hydroxylase (CYP21) deficiency congenital adrenal hyperplasia; [2] 11-beta-hydroxylase (CYP11B1) deficiency; congenital adrenal hyperplasia; [3] aldosterone synthase (CYP11B2) deficiency; [4] 3 beta-hydroxysteroid dehydrogenase (3b-HSD) deficiency; congenital adrenal hyperplasia; [5a] 17-hydroxylase deficiency; congenital adrenal hyperplasia; [5b] 17,20 desmolase deficiency 17-hydroxylase and 17,20 desmolase; both encoded by (CYP17).

Disorders of glucocorticoid deficiency (adrenal insufficiency): laboratory testing

When adrenal insufficiency is clinically suspected, the ability of the adrenal gland to secrete cortisol is evaluated by the ACTH (cosyntropin and short synacthen) stimulation test. This test can be performed at any time of the day and involves administration of ACTH after a baseline cortisol level, followed by levels at 30 and 60 minutes. A normal response would be a rise in cortisol greater than 20 μg/dL [9]. Failure to mount an adequate cortisol response indicates primary adrenal insufficiency. ACTH levels should also be measured (prior to stimulation test) to determine whether the defect is at the level of adrenal gland (will be elevated) or "higher" (if due to pituitary disorder, ACTH will be depressed). Additional tests include adrenal autoantibody testing and imaging to determine if there is anatomic disease.

The diagnosis of Addison disease in the ICU setting is controversial yet failure to treat true Addison disease can be fatal especially in the setting of illness. However, there are adverse consequences associated with its treatment as pharmacologic doses of glucocorticoids can result in immunosuppression, hyperglycemia, impaired wound healing, and elevated blood pressure. Adrenal cortical deficiency in the critical care setting has been defined as an increase in the cortisol of <9 μg/dL in response to 250 μg of cosyntropin or a random cortisol of <10 μg/dL.

Adrenal androgens

Adrenal androgens have their greatest effect in women where they are responsible for the development of axillary and pubic hair. In men, the effects of adrenal androgens are minor, because testosterone is primarily produced in the testes with much smaller amounts produced in the adrenals. ACTH stimulates the synthesis and secretion of DHEA and androstenedione; however, physiological control of adrenal androgen synthesis and release is unclear.

Disorders of adrenal androgen excess (hyperandrogenism)

Pathologic increases in adrenal androgens can cause premature adrenarche or hirsutism/virilization in women (with oligomenorrhea, amenorrhea, and infertility as additional possible consequences). Hyperandrogenism is commonly due to CAH but can occur with certain tumors of reproductive system. Androgen producing adenomas and carcinomas are rare. Exogenous adrenal androgens (or other androgens, such as testosterone) in pharmacologic doses produce many adverse health outcomes such as infertility (from gonadotropin suppression), psychological disease (e.g., aggressive behavior), and an increased risk of cardiovascular disease from elevated blood pressure, elevations in low-density lipoprotein-cholesterol, and depressions in high-density lipoprotein-cholesterol.

Premature adrenarche occurs when adrenal androgen production increases earlier than normal (7−8 years of age), leading to axillary and pubic hair in the absence of any other evidence of puberty (e.g., girls do not estrogenize and boys do not display increased size of the penis or testes). Hyperandrogenism in utero in female fetuses causes virilization of the external genitalia resulting in clitoromegaly and labial fusion. The most common cause is CAH in the form of 21-OH deficiency or 11B-OH deficiency.

Such types of CAH result in overproduction of adrenal androgens (DHEA and androstenedione). In very severe cases of CAH, the fetal fetus will be virilized to the point where there is a penile urethra and what appears to be a scrotum with bilaterally undescended testes. In the older literature, CAH was referred to as the "adrenogenital syndrome" [12].

In women of reproductive age, the most common cause of hyperandrogenism is polycystic ovary syndrome (PCOS). Women with PCOS have many features of the metabolic syndrome: centripetal obesity, hirsutism, irregular menses, amenorrhea, or infertility. In this disorder, the ovary secretes excessive testosterone. To measure small, but significant increases in testosterone in women, testosterone should be measured by mass spectrometry (MS) due to a greater analytic sensitivity than testosterone immunoassays. Using this value and a measurement of sex-hormone binding globulin, free testosterone can be calculated. MS assays for free testosterone are available and present a more accurate measurement than the calculated values.

Disorders of adrenal androgen excess (hyperandrogenism) laboratory testing

Excesses of DHEA, dehydroepiandrosterone sulfate (DHEA-S), or androstenedione are investigated by measuring the fasting levels of these hormones in plasma. CAH should be investigated by measuring 17-OHP. Fig. 41.3 illustrates the metabolic blocks in 21-OH (CYP21A2) deficiency and 11beta-hydroxylase deficiency.

Urinary 17-ketosteroids (17-KS) are the metabolites of DHEA and androstenedione and can be measured in a 24-hour urine collection as an integrated measure of adrenal androgen production. In 21-OH-deficient and 11-OH-deficient CAH, 17-KS are elevated and replacement therapy with cortisol (or another appropriately dosed glucocorticoid) will suppress 17-KS to normal. However, affected patients are more easily managed through present-day measurements of serum or plasma 17-OHP or androstenedione. Prior to the development of plasma 17-OHP or androstenedione measurements, 24-hour urinary 17-ketogenic steroids (metabolites of 17-OHP, 11-desoxycortisol, and cortisol) were measured in addition to 17-KS to diagnosis and manage 21-OH (CYP21A2)-deficient CAH.

There is a late-onset form of 21-OH-deficient CAH, where the enzyme defect is relatively mild. There is no virilization in utero of the female fetus. However, at the time of puberty, hirsutism, virilization, and menstrual irregularities develop. Fasting 17-OHP may only be marginally elevated; however, the response of 17-OHP to cosyntropin injection is excessive and identify this form of 21-OH-deficient CAH. Individuals who carry mild mutations in their 21-OH genes are asymptomatic, yet have excessive 17-OHP responses to cosyntropin injection and are classified as having cryptic 21-OH-deficient CAH.

Less common forms of CAH are 3-beta-hydroxysteroid dehydrogenase (3b-HSD) deficiency and CYP17 deficiency. In 3b-HSD deficiency, there is a defective conversion of ^{5}delta to ^{4}delta steroids (e.g., defects in the following conversions: pregnenolone → progesterone; 17-hdroxypregnenolone → 17-OHP; and DHEA → androstenedione) with deficiencies of aldosterone and cortisol. Affected males are incompletely virilized in utero (because there is a block in testosterone

biosynthesis), whereas females are virilized in utero (because of the accumulation of DHEA). Diagnosis is made by examining the increased ratio of ^{5}delta to ^{4}delta steroids. In CYP17 deficiency, there is defective formation of sex steroids and cortisol. Males are undervirilized (or phenotypically female), whereas females fail to feminize at the time of puberty and suffer from primary amenorrhea. The least common form of CAH results from a loss-of-function mutation in steroidogenic acute regulatory protein. This results in deficiencies of all adrenal and gonadal steroids that are usually fatal. The males will display ambiguous genitalia or a frankly female phenotype. In the absence of CAH with elevations in DHEA, DHEA-S, androstenedione, and/or (far less commonly) testosterone, an androgen-secreting adrenal adenoma or carcinoma must be sought [13].

Disorders of adrenal androgen deficiency

Although adrenal androgen deficiency can occur with Addison disease, there are no apparent adverse consequences.

Disorders of adrenal androgen deficiency laboratory testing

There are no readily recognized adverse consequences of isolated adrenal androgen deficiency after birth, and measurements of DHEA, DHEA-S, or androstenedione are not considered part of its diagnostic workup. On the other hand, biosynthetic defects in the overall production of adrenal androgens, for example, 17-hydroxylase/17,20-desmolase (CYP17) deficiency and 3b-HSD, impair testosterone, and estradiol synthesis, will have significant clinical impacts.

In 17-hydroxylase/17,20 desmolase deficiency, males are ambiguous or phenotypically female. There is cortisol deficiency but mineralocorticoids are not deficient, and elevated 11-desoxycorticosterone (DOC) leads to hypokalemia/alkalotic hypertension in childhood. In contrast, females with 17-hydroxylase/17,20 desmolase deficiency do not have maldevelopment of the internal or external genitalia; however, hypokalemia/alkalotic hypertension can develop. At puberty, estrogen deficiency occurs with pubertal failure (e.g., a lack of estrogenization) manifested as primary amenorrhea. Deficiency of 3b-HSD causes sexual ambiguity in males and females. In males, there is insufficient androgen to virilize fully the male fetus. On the other hand, elevated DHEA levels virilize the female fetus. With 3b-HSD deficiency, there is deficient synthesis of cortisol and aldosterone. Adrenal androgen concentrations decline in men and women over time; however, DHEA or androstenedione are not accepted treatments and their therapeutic use is highly controversial.

The adrenal medulla

Located in the center of the adrenal gland, the medulla represents ~10% of the total weight of each adrenal (i.e., 0.4−0.6 g). Despite the small mass of the adrenal medulla, the catecholamine products of the medulla are critically important for survival. Adrenal medullary cells stain brown with chromium salts secondary to catecholamine oxidation to melanin; consequently, they are termed chromaffin cells or pheochromocytes.

Chromaffin cells are a type of modified sympathetic postganglionic neuron that lacks postganglionic processes secreting either epinephrine or norepinephrine. The synthesis of catecholamines is outlined in Fig. 41.4. The medulla secretes epinephrine and lesser amounts of norepinephrine and dopamine. In contrast, peripheral sympathetic nerves only secrete norepinephrine. Epinephrine and norepinephrine are stored in secretory granules along with chromogranin A, ATP, proopiomelanocorticotropin, and adrenomedulin. Interestingly, the systemic effects of catecholamines result from epinephrine secretion, while circulating norepinephrine levels are not sufficiently high to produce systemic effects and the physiologic role of adrenal dopamine is unclear.

The adrenal medulla is an important element of the sympathetic nervous system. The sympathetic nervous system is key to mammals' successful responses to threats that require the "fight and flight" response. Catecholamines raise blood glucose and blood pressure to prepare for the "fight" (e.g., defensive actions) and/or the "flight" (escape). Catecholamines stimulate the liver to increase hepatic glucose output by stimulating glycogenolysis and increased amino acid and lactate uptake to increase gluconeogenesis. Lactate is a product of anaerobic glycolysis from exercising muscle. Rising free fatty acid concentrations provide an alternative energy source to glucose, which will act to maintain or increase blood glucose concentrations. Epinephrine effects on the cardiovascular system include vasoconstriction in the renal, intestinal, and cutaneous circulation, vasodilatation in skeletal muscle and the coronary circulation, venous constriction, increased cardiac inotropism, increased cardiac conduction velocity, and tachycardia. Bronchodilatation and relaxation of intestinal smooth muscle and uterine smooth muscle are other effects of catecholamines.

Medullary epinephrine secretion is controlled by the autonomic nervous system. Splanchic nerve preganglionic sympathetic fibers innervate chromaffin cells via the actions of secreted acetylcholine binding to ganglionic-type nicotinic receptors of the chromaffin cells. This

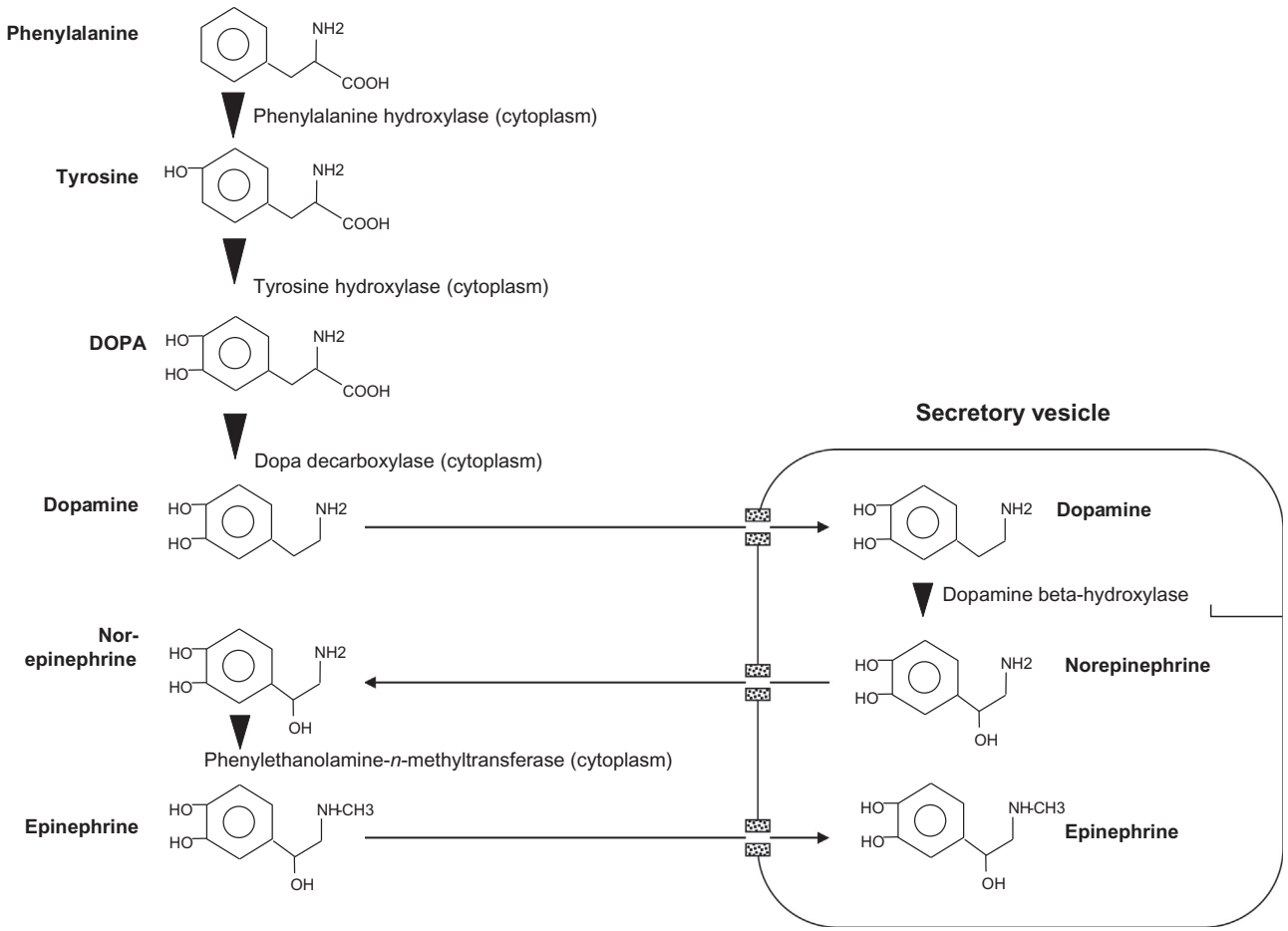


FIGURE 41.4 Catecholamine synthetic pathway.

process triggers the release of chromaffin cell secretory granules, liberating epinephrine and norepinephrine.

Disorders of catecholamine excess

Excessive secretion of catecholamines occurs as a consequence of adrenal medullary chromaffin-cell tumors. Pheochromocytoma is the classic medullary tumor. Clinically, patients experience severe, episodic or persistent hypertension, tachycardia, sweating, pallor, headache, and a "feeling of doom" during times of catecholamine release. Other findings encompass nervousness, nausea, and vomiting, hypertensive retinopathy, fasting hyperglycemia, orthostatic hypotension, and decreased weight. Hypertension can precipitate myocardial infarction or stroke. Pheochromocytoma is most frequently diagnosed in the fourth or fifth decade.

Ten percent of pheochromocytomas are malignant, ~10% are bilateral, ~10% occur in children (where they are more commonly familial and bilateral), and ~10% are located outside of the adrenal medulla (e.g., in the organ of Zuckerkandl). Extra adrenal chromaffin-cell tumors are termed "extra-adrenal" paragangliomas (a pheochromocytoma is an intraadrenal paraganglioma). Histologic examination of pheochromocytomas exhibits cellular features that range from normal to severe atypia and malignancy can only be established by finding metastases. Previously, it was believed that ~10% of pheochromocytomas were familial, but more recent studies report that ~25% of pheochromocytomas are familial. Familial pheochromocytoma can occur as a dominant disorder or as part of von Hippel–Landua type 2 syndrome, Sturge–Weber syndrome, von Recklinghausen neurofibromatosis type 1, tuberous sclerosis, or multiple endocrine neoplasia type 2 (MEN 2).

There are two types of MEN: MEN 1 and MEN 2. MEN 2 has two subtypes: 2A and 2B (formerly known as MEN 3). MEN 1 results from *MENIN* mutations that are inherited in a autosomal dominant fashion and are characterized by pituitary (most commonly prolactinomas), parathyroid (hyperplasia), and pancreatic islet (most commonly gastrinomas) tumors. Tumorogenesis occurs following the loss of heterozygosity in affected tissues that produce hemizygosity for the mutated and dysfunctional

tumor-suppressor gene *MENIN*. MEN 2 results from gain-of-function mutations in the RET protooncogene that encodes a cell-surface receptor of the tyrosine kinase family. Approximately 50% of affected individuals will develop a pheochromocytoma, whereas 100% of affected individuals will develop medullary thyroid carcinoma (MTC). These tumors can develop in childhood with MTC even developing in infancy. See Chapter 44, Tumor markers, of this book for a further discussion of MEN 1, MEN 2, and the distinctive characteristics of MEN 2A versus 2B.

Other tumors of neural crest origin include ganglioneuromas, ganglioneuroblastomas, and neuroblastomas. Neuroblastomas occur in infants or children and may be present at birth. It presents as either an abdominal mass in an infant or child or with widespread metastases. While these tumors may produce catecholamines, there are insufficient elevations of catecholamines to produce clinical symptoms. Nevertheless, catecholamine metabolites measured in timed urine samples can serve as tumor markers (see below).

Disorders of catecholamine excess laboratory testing

Historically, testing was carried out by performing catecholamine measurements in plasma or urine. Test descriptions may vary among reference laboratories, so it is important to clarify which catecholamines are being measured in each assay. Similar test names with different components include "plasma catecholamines" (dopamine plus norepinephrine plus epinephrine), "total" plasma catecholamines (norepinephrine and epinephrine), and "fractionated plasma catecholamines" (specific individual measurements of dopamine, norepinephrine, and epinephrine).

Catecholamine testing is subject to many preanalytic interferences including posture during blood collection, age, placement of samples postcollection on ice, and painful phlebotomy procedure. Use of a butterfly needle with heparin flush to maintain patency with a 20-minute "recovery period" from the needle's insertion is recommended. During this time, the elevated catecholamine levels related to phlebotomy will return to baseline. Biochemical testing for pheochromocytoma is hindered by the fact that catecholamine release is proportional to symptomology, such that if the patient is asymptomatic, catecholamine levels might be falsely depressed.

With the advent of "plasma-free" (nonsulfated) metanephrine measurements, diagnosis has become easier. Plasma-free metanephrines are continuously produced by pheochromocytoma cells, making them less subject to the episodicity of catecholamine release. Plasma-free metanephrine testing is the plasma assay of choice for the diagnosis of pheochromocytoma with a sensitivity and specificity of 97% and 93%, respectively. It is recommended that the patient be supine for up to 30 minutes prior to collection to minimize false-positive results. Twenty-four-hour urine collections can also be performed in search of catecholamine excess. Testing options include "fractionated urine catecholamines," which report specific measurements of dopamine, norepinephrine, and epinephrine per 24 hour [or per gram of creatinine if a random (spot) sample is submitted]. Total urine catecholamines (i.e., norepinephrine plus epinephrine) may also be reported. To preserve urine samples, acid is added to the container (e.g., 20−25 mL of 6 N HCl or 15 g of boric acid).

Initial metabolites of the catecholamines are normetanephrine and metanephrine (collectively referred to as "metanephrines"). Urine metanephrine and normetanephrine can also be individually measured (e.g., fractionated urine metanephrines). Elevated excretion of 24-hour urine metanephrines is superior to the diagnosis of pheochromocytoma compared with other measures of urine catecholamine excretion. However, plasma catecholamine measurements are highly labile and may be elevated to pain and stress leading to false-positive results (i.e., elevated catecholamines in the absence of a pheochromocytoma).

A dynamic test for catecholamine excess is the clonidine suppression test. Clonidine is a centrally acting alpha-2-adrenergic agonist. Blood pressure and plasma catecholamines are drawn at baseline (i.e., 1 hour after insertion of an intravenous line with the patient supine). Three hours after oral administration of clonidine (0.3 mg/70 kg), blood pressure and plasma catecholamines are again measured. Some sources also recommend measuring blood pressure and plasma catecholamines 1 and 2 hours after administration of the clonidine. In normal individuals, blood pressure and plasma catecholamines decline. In pheochromocytoma, plasma catecholamines remain unchanged or rise in concentration. This test is only informative when basal plasma catecholamines increase. Because of the possibility of acute hypertensive crisis, provocative testing with tyramine, glucagon, metoclopramide, or histamine is not advised in the evaluation of suspected pheochromocytoma.

For plasma or urine testing, it is advised that subjects avoid strenuous exercise, tobacco, tea, coffee, and alcohol prior to testing. For urine testing, common drugs used to treat hypertension (e.g., alpha or beta blockers, ACE inhibitors, calcium channel blockers, and diuretics) produce little or no interferences. In the case of urine metanephrine testing, it is recommended that patients be off medications for 3 days prior to testing. Alpha blockers and alpha agonists are not to be taken for 18−24 hours prior the beginning of the urine collection.

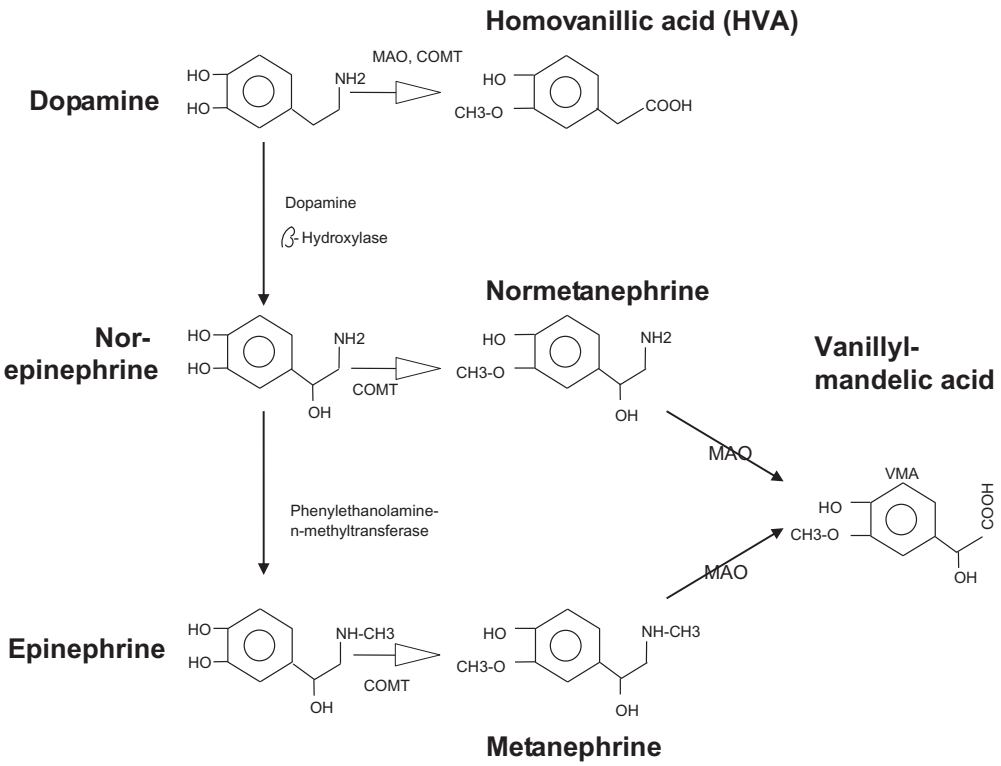


FIGURE 41.5 Catecholamine metabolites.

Homovanillic acid (HVA) and vanillylmandelic acid (VMA) are the metabolites of dopamine and norepinephrine, respectively (Fig. 41.5). Serum and urine levels of HVA and VMA are typically measured to diagnose and monitor children with neuroblastoma, but may also be utilized in pheochromocytoma. Typically, both are assayed using high-performance liquid chromatography [14,15].

Catecholamine deficiency

Inadequate counter regulatory epinephrine and/or glucagon responses to hypoglycemia have been reported in individuals after 5–10 years of diabetes mellitus. This impairs the diabetic patient's ability to recover spontaneously from hypoglycemia that may have resulted from a missed meal, excessive exercise, and/or overdosing of insulin or an oral hypoglycemic agent. While hyperglycemia and its complications are a chronic challenge to patients with diabetes, hypoglycemia is usually an acute problem that can be life-threatening.

References

[1] D.L. Fisher, Overview of the renin-angiotensin system, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA (accessed February 2016).

[2] M. Stowasser, Assays of the renin-angiotensin-aldosterone system in adrenal disease, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA (accessed February 2016).

[3] M. Boscaro, V. Ronconi, F. Turchi, G. Giacchetti, Diagnosis and management of primary aldosteronism, Curr. Opin. Endocrinol. Diabetes Obes. 4 (2008) 332–338.

[4] R.M. Carey, S.H. Padia, Primary mineralocorticoid excess disorders and hypertension, in: J.L. Jameson, L.J. De Groot, D.M. de Kretser, L.C. Giudice, A.B. Grossman, S. Melmed, J.T. Potts, G.C. Weir (Eds.), Endocrinology: Adult and Pediatric, Saunders, Philadelphia, PA, 2016, pp. 1871–1892. Downloaded from Clinicalkey.com 3/2016.

[5] J.W. Funder, R.M. Carey, F. Mantero, M.H. Murad, M. Reincke, H. Shibata, et al., The management of primary aldosteronism: case detection, diagnosis,and treatment: an Endocrine Society Clinical Practice Guideline, J. Clin. Endocrinol. Metab. (2016) [Epub ahead of print].

[6] A. Ganguly, Distinguishing unilateral aldosteronoma from bilateral disease, J. Clin. Endocrinol. Metab. 8 (2001) 4004–4005.

[7] W.E. Winter, N.S. Harris, Laboratory evaluation of endocrine hypertension, in: W.E. Winter, L. Sokoll, I. Jialal (Eds.), Handbook of Diagnostic Endocrinology, second ed., American Association of Clinical Chemistry Inc., 2008, pp. 175–210.

[8] J.T. Li, S.G. Shu, C.S. Chi, Aldosterone-secreting adrenal cortical adenoma in an 11-year-old child and collective review of the literature, Eur. J. Pediatr. 10 (1994) 715–717.

[9] T.J. McKenzie, J.B. Lillegard, W.F. Young Jr., G.B. Thompson, Aldosteronomas—state of the art, Surg. Clin. North. Am. 5 (2009) 1241–1253.

[10] D.V. Milford, Investigation of hypertension and the recognition of monogenic hypertension, Arch. Dis. Child. 5 (1999) 452–455.
[11] P. Mulatero, C. Bertello, A. Verhovez, D. Rossato, G. Giraudo, G. Mengozzi, et al., Differential diagnosis of primary aldosteronism subtypes, Curr. Hypertens. Rep. 3 (2009) 217–223.
[12] M. Rehan, J.E. Raizman, E. Cavalier, A.C. Don-Wauchope, D.T. Holmes, Laboratory challenges in primary aldosteronism screening and diagnosis, Clin. Biochem. 48 (2015) 377–387.
[13] G.M. Rich, S. Ulick, S. Cook, J.Z. Wang, R.P. Lifton, R.G. Dluhy, Glucocorticoid-remediable aldosteronism in a large kindred: clinical spectrum and diagnosis using a characteristic biochemical phenotype, Ann. Intern. Med. 10 (1992) 813–820.
[14] G.P. Rossi, T.M. Seccia, A.C. Pessina, Primary aldosteronism - part I: prevalence, screening, and selection of cases for adrenal vein sampling, J. Nephrol. 4 (2008) 447–454.
[15] G.P. Rossi, T.M. Seccia, A.C. Pessina, Primary aldosteronism: part II: subtype differentiation and treatment, J. Nephrol. 4 (2008) 455–462.

Further readings

M. Andrioli, G.F. Pecori, M.M. De, A. Cattaneo, C. Carzaniga, F. Cavagnini, Differential diagnosis of ACTH-dependent hypercortisolism: imaging versus laboratory, Pituitary 4 (2009) 294–296.
Z. Antal, P. Zhou, Congenital adrenal hyperplasia: diagnosis, evaluation, and management, Pediatr. Rev. 7 (2009) e49–e57.
R.J. Auchus, W. Arlt, Approach to the patient: the adult with congenital adrenal hyperplasia, J. Clin. Endocrinol. Metab. 98 (2013) 2645–2655.
I. Bancos, S. Hahner, J. Tomlinson, W. Arlt, Diagnosis and management of adrenal insufficiency. Lancet Diabetes Endocrinol. 2015;3:216-26 Bouillon R, Acute adrenal insufficiency, Endocrinol. Metab. Clin. North. Am. 4 (2006) 767–775.
A. Barthel, H.S. Willenberg, M. Gruber, S.R. Bornstein, Adrenal Insufficiency, in: J.L. Jameson, L.J. De Groot, D.M. de Kretser, L.C. Giudice, A.B. Grossman, S. Melmed, J.T. Potts, G.C. Weir (Eds.), Endocrinology: Adult and Pediatric, Saunders, Philadelphia, PA, 2016, pp. 1902–1927. Downloaded from Clinicalkey.com 3/2016.
M. Boscaro, G. Arnaldi, Approach to the patient with possible Cushing's syndrome, J. Clin. Endocrinol. Metab. 9 (2009) 3121–3131.
E. Charmandari, N.C. Nicolaides, G.P. Chrousos, Adrenal insufficiency, Lancet. 21 (2014) 2152–2167.
G.P. Chrousos, Adrenal hyperandrogenism, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA (accessed February 2016).
G. Eisenhofer, G. Siegert, J. Kotzerke, S.R. Bornstein, K. Pacak, Current progress and future challenges in the biochemical diagnosis and treatment of pheochromocytomas and paragangliomas, Horm. Metab. Res. 40 (5) (2008) 329–337.
M.B. Elamin, M.H. Murad, R. Mullan, D. Erickson, K. Harris, S. Nadeem, et al., Accuracy of diagnostic tests for Cushing's syndrome: a systematic review and metaanalyses, J. Clin. Endocrinol. Metab. 5 (2008) 1553–1562.
M.M. Erichsen, K. Lovas, B. Skinningsrud, A.B. Wolff, D.E. Undlien, J. Svartberg, et al., Clinical, immunological, and genetic features of autoimmune primary adrenal insufficiency: observations from a Norwegian registry, J. Clin. Endocrinol. Metab. 12 (2009) 4882–4890.
L. Fishbein, Pheochromocytoma and paraganglioma: genetics, diagnosis, and treatment, Hematol. Oncol. Clin. North. Am. 30 (2016) 135–150.
C. Garcia, B.M. Biller, A. Klibanski, The role of the clinical laboratory in the diagnosis of Cushing's syndrome, Am. J. Clin. Pathol. (2003) S38–S45.
I.A. Hughes, Congenital adrenal hyperplasia: a lifelong disorder, Horm. Res. (2007) 84–89.
E. Husebye, K. Lovas, Pathogenesis of primary adrenal insufficiency, Best Pract. Res. Clin. Endocrinol. Metab. 2 (2009) 147–157.
S. Jehle, J.E. Walsh, P.U. Freda, K.D. Post, Selective use of bilateral inferior petrosal sinus sampling in patients with adrenocorticotropin-dependent Cushing's syndrome prior to transsphenoidal surgery, J. Clin. Endocrinol. Metab. 12 (2008) 4624–4632.
A. Juszczak, D.G. Morris, A.B. Grossman, L.K. Nieman, Cushing's syndrome, in: J.L. Jameson, L.J. De Groot, D.M. de Kretser, L.C. Giudice, A.B. Grossman, S. Melmed, J.T. Potts, G.C. Weir (Eds.), Endocrinology: Adult and Pediatric, Saunders, Philadelphia, PA, 2016, pp. 227–255. Downloaded from Clinicalkey.com 3/2016.
A. Lacroix, Dexamethasone suppression tests, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA (accessed February 2016).
C.L. Lin, T.J. Wu, D.A. Machacek, N.S. Jiang, P.C. Kao, Urinary free cortisol and cortisone determined by high performance liquid chromatography in the diagnosis of Cushing's syndrome, J. Clin. Endocrinol. Metab. 1 (1997) 151–155.
P.E. Marik, S.M. Pastores, D. Annane, G.U. Meduri, C.L. Sprung, W. Arlt, et al., American College of Critical Care Medicine. Recommendations for the diagnosis and management of corticosteroid insufficiency in critically ill adult patients: consensus statements from an international task force by the American College of Critical Care Medicine, Crit. Care Med. 36 (2008) 1937–1949.
D.P. Merke, Genetics and clinical presentation of classic congenital adrenal hyperplasia due to 21-hydroxylase deficiency, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA (accessed February 2016).
L.K. Nieman, Causes and pathophysiology of Cushing's syndrome, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA (accessed February 2016).
L.K. Nieman, Cushing's syndrome: update on signs, symptoms and biochemical screening, Eur. J. Endocrinol. 173 (2015) 33–38.
L.K. Nieman, Diagnosis of adrenal insufficiency, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA (accessed February 2016).
L.K. Nieman, Establishing the diagnosis of Cushing's syndrome, in: T. W. Post (Ed.), UpToDate, UpToDate, Waltham, MA (accessed February 2016).
L.K. Nieman Evaluation of the response to ACTH in adrenal insufficiency, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA (accessed February 2016).
L.K. Nieman, Measurement of cortisol in serum and saliva, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA (accessed February 2016).
L.K. Nieman, Measurement of urinary excretion of endogenous and exogenous glucocorticoids, in: T.W. Post (Ed.), UpToDate, UpToDate, Waltham, MA (accessed February 2016).
L.K. Nieman, B.M. Biller, J.W. Findling, J. Newell-Price, M.O. Savage, P.M. Stewart, et al., The diagnosis of Cushing's syndrome: an Endocrine Society Clinical Practice Guideline, J. Clin. Endocrinol. Metab. 5 (2008) 1526–1540.

L.K. Nieman, Dynamic evaluation of adrenal hypofunction, J. Endocrinol. Invest. 7 (Suppl) (2003) 74–82.
E.H. Oldfield, J.L. Doppman, L.K. Nieman, G.P. Chrousos, D.L. Miller, D.A. Katz, et al., Petrosal sinus sampling with and without corticotropin-releasing hormone for the differential diagnosis of Cushing's syndrome, N. Engl. J. Med. 13 (1991) 897–905.
K. Pacak, H.J.L.M. Timmers, Eisenhofer, Pheochromocytoma, in: J.L. Jameson, L.J. De Groot, D.M. de Kretser, L.C. Giudice, A.B. Grossman, S. Melmed, J.T. Potts, G.C. Weir (Eds.), Endocrinology: Adult and Pediatric, Saunders, Philadelphia, PA, 2016, pp. 1902–1927. Downloaded from Clinicalkey.com 3/2016.
D.A. Papanicolaou, J.A. Yanovski, G.B. Cutler Jr., G.P. Chrousos, L.K. Nieman, A single midnight serum cortisol measurement distinguishes Cushing's syndrome from pseudo-Cushing's states, J. Clin. Endocrinol. Metab. 4 (1998) 1163–1167.
G.F. Pecori, Recent challenges in the diagnosis of Cushing's syndrome, Horm. Res. (2009) 123–127.
H. Raff, Cushing syndrome: update on testing, Endocrinol. Metab. Clin. North. Am. 44 (2015) 43–50.
R.J. Singh, Advances in metanephrine testing for the diagnosis of pheochromocytoma, Clin. Lab. Med. 24 (1) (2004) 85–103.
P.M. Stewart, M.O. Quinkler, Mineralocorticoid deficiency, in: J.L. Jameson, L.J. De Groot, D.M. de Kretser, L.C. Giudice, A.B. Grossman, S. Melmed, J.T. Potts, G.C. Weir (Eds.), Endocrinology: Adult and Pediatric, Saunders, Philadelphia, PA, 2016, pp. 1892–1901. Downloaded from Clinicalkey.com on 3/2016.
M. Stowasser, Update in primary aldosteronism, J. Clin. Endocrinol. Metab. 100 (2015) 1–10.
W.E. Winter, N.S. Harris, Laboratory approaches to diseases of the adrenal cortex and adrenal medulla, in: W.E. Winter, L. Sokoll, I. Jialal (Eds.), Handbook of Diagnostic Endocrinology, second ed., American Association of Clinical Chemistry Inc., 2008, pp. 75–138.
M. Yau, A. Khattab, C. Pina, T. Yuen, H.F.L. Meyer-Bahlburg, N. Maria, Defects in Adrenal Steroidogenesis, in: J.L. Jameson, L.J. De Groot, D.M. de Kretser, L.C. Giudice, A.B. Grossman, S. Melmed, J.T. Potts, G.C. Weir (Eds.), Endocrinology: Adult and Pediatric, Saunders, Philadelphia, PA, 2016, pp. 1811–1832. Downloaded from Clinicalkey.com 3/2016.

Self-assessment questions

1. A lean, nondiabetic 45-year-old hypertensive male who is on no therapies has a potassium concentration of 3.0 meq/L or lower on multiple occasions (reference interval: 3.5–5.0 meq/L). Which one of the following tests is likely to be of the greatest diagnostic assistance?
   a. Cortisol following 1 mg of dexamethasone given the previous evening
   b. Antidiuretic hormone measurement (vasopressin)
   c. Renin and aldosterone measurements
   d. Plasma-free metanephrine measurement
   e. 24-hour urinary HVA and VMA measurements
2. An obese, diabetic 58-year-old hypertensive male who is on no diuretics has a potassium concentration of 3.0 meq/L or lower on multiple occasions (reference interval: 3.5–5.0 meq/L). He has problems with poor wound healing and infections. He has had a collapsed vertebra in the past. Which one of the following tests is likely to be of the greatest diagnostic assistance?
   a. Cortisol following 1 mg of dexamethasone given the previous evening
   b. Antidiuretic hormone measurement (vasopressin)
   c. Renin and aldosterone measurements
   d. Plasma-free metanephrine measurement
   e. 24-hour urinary HVA and VMA measurements
3. Which one of the following assays is least helpful in diagnosing Cushing syndrome?
   a. Cortisol following 1 mg of dexamethasone given the previous evening
   b. Cortisol drawn at midnight
   c. 24-hour urine-free cortisol
   d. a.m. cortisol
4. Which one of the following plasma assays is most helpful in diagnosing pheochromocytoma?
   a. Total catecholamines
   b. Plasma free metanephrines
   c. Epinephrine
   d. Norepinephrine
   e. Dopamine
5. A hyperpigmented patient presents with hypotension, hypoglycemia, hyponatremia, hyperkalemia, and acidosis. What is the most likely diagnosis?
   a. Cortisol resistance
   b. Addison disease
   c. Isolated mineralocorticoid deficiency
   d. ACTH deficiency
   e. Apparent mineralocorticoid excess
6. An infant is born with ambiguous genitalia. Which one of the following tests would be most helpful diagnostically?
   a. Cortisol
   b. Pregnenolone
   c. 24-hour urinary 17-hydroxycorticosterone
   d. Testosterone
   e. 17-hydroxyprogesterone
7. An infant has a palpable abdominal mass. The blood pressure is normal. Which one of the following tests, if normal, would be most helpful in excluding adrenal pathology?
   a. 24-hour urinary HVA and VMA
   b. Aldosterone after salt loading
   c. 24-hour urinary metanephrines
   d. Cortisol after cosyntropin injection
   e. Plasma-free metanephrines
8. Which one of the following tests is most helpful in differentiating an aldosteronoma from bilateral idiopathic adrenal hyperplasia?
   a. Measurement of ACTH in the inferior petrosal venous sinuses and inferior vena cava at baseline and following CRH administration

b. A cortisol measurement following repeated administrations of cosyntropin
c. Failure of aldosterone to increase when the subject is upright
d. Measurement of the aldosterone to renin ratio in a peripheral blood sample
e. An MRI or CT of the pituitary gland

9. A lean, nondiabetic patient is hypertensive, hypokalemic, and alkalotic. They are not treated with any medications. The renin and aldosterone are both decreased. Which of the following disorders is least likely?
a. Glucocorticoid remediable hyperaldosteronism
b. Apparent mineralocorticoid excess
c. Inhibition of 11-beta-hydroxysteroid dehydrogenase-2
d. Gain-of-function mutation in ENaC
e. Gain-of-function mutation in the mineralocorticoid receptor

10. A hyperpigmented patient presents with hypotension, hypoglycemia, hyponatremia, hyperkalemia, and acidosis. What is the most likely cause of the patient's underlying disorder?
a. An anterior pituitary corticotropinoma
b. Autoimmune destruction of the adrenal cortex
c. Administration of high-dose glucocorticoids
d. Failure of HSD11B2 to convert cortisol into cortisone
e. Adrenoleukodystrophy

Answers

1. c
2. a
3. d
4. b
5. b
6. e
7. a
8. c
9. a
10. b

Chapter 42

Laboratory testing in pregnancy

Alison Woodworth[1] **and Christopher R. McCudden**[2]

[1]*Department of Pathology and Laboratory Medicine, University of Kentucky Medical Center, Lexington, KY, United States,* [2]*Eastern Ontario Regional Laboratory Association, The Ottawa Hospital Department of Pathology and Laboratory Medicine, University of Ottawa, ON, Canada*

Learning objectives

After reviewing this chapter, the reader will be able to:

- Explain the changes in physiology and laboratory values associated with normal and abnormal pregnancies.
- Explain the differences among human chorionic gonadotropin assays, their utility, and limitations.
- Describe screening and diagnostic tests for fetal and maternal complications in pregnancy and their limitations and how these are interpreted.

Normal pregnancy

Pregnancy is a healthy state lasting ~40 weeks from the first day of the last menstrual period (LMP) to the expected date of confinement or estimated delivery date. The 40-week time frame is divided into three trimesters of 13 weeks; the first trimester begins with the first day of the LMP. A term pregnancy is any that ends in delivery of the infant between 37 and 42 weeks of gestation. New guidance suggests that a full-term pregnancy is any delivery between 39 and 40 weeks and 6 days, while early term is any pregnancy between 37 and 38 weeks and 6 days, and late and postterm pregnancies are those with deliveries during the 41st and beyond 42nd weeks, respectively. Evidence suggests that full-term pregnancies are associated with the best overall outcome for the infant.

The placenta, umbilical cord, and amniotic fluid (AF) are critical for maintaining pregnancy. The placenta separates the maternal and fetal circulation systems while providing nourishment and eliminating fetal waste. Concentration gradients regulate transport of proteins and small molecules across the placenta. With the exception of maternal IgG, which crosses via receptor-mediated endocytosis, large proteins do not cross the placental membranes. The placenta also produces hormones critical for maintaining pregnancy, such as human chorionic gonadotropin (hCG), human placental lactogen, progesterone, and estrogen. During pregnancy, the fetus is protected and suspended by AF. The volume and composition of AF change over the course of pregnancy through bidirectional flow of molecules along the maternal, fetal, and placental membranes. Analytes detected in AF are used to assess abnormalities in pregnancy including fetal lung immaturity, congenital anomalies, and hemolytic disease.

Physiological changes during pregnancy

Physiological changes in pregnancy affect the hemodynamic, cardiovascular, renal, hepatic, endocrine, and respiratory systems and their associated laboratory tests. Assessment of pregnancy abnormalities requires an understanding of these normal changes. The changes in common analytes that occur during pregnancy, expressed as a percentage of expected nonpregnant reference intervals, are listed in Table 42.1.

Significant hematological changes occur during pregnancy. Maternal blood volume increases by ~45%; of this about 20% is due to red blood cell (RBC) mass and 80% to plasma volume. As a result, hematocrit, hemoglobin, and erythrocyte count are decreased, a combination commonly referred to as "physiological anemia of pregnancy." Pregnancy is a hypercoagulable state associated with increases in several clotting factors and significant risk of deep vein thrombosis. In addition, maternal cardiac output is increased by 30%−40%, resulting in increased blood pressure and flow to organs. Increased blood flow to the kidneys increases glomerular filtration. Renal function tests reveal an elevated eGFR with corresponding decreases in creatinine and BUN.

Pregnancy is marked by significant changes in the endocrine system that are required to maintain gestation, promote labor and delivery of term infants, and facilitate lactation. Increases in the pregnancy-maintaining hormones hCG, progesterone, and estrogen occur early in pregnancy and persist throughout. A combination of

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00042-9

TABLE 42.1 Changes in normal laboratory values during pregnancy.

Analyte	Percentage of normal pregnancy mean		
	First trimester	Second trimester	Third trimester
Hemodynamic			
Erythrocyte count	94	86	93
Erythropoietin	100	114	156
Factor VII	118	160	189
Factor X	114	128	141
Hemoglobin	95	89	93
Hematocrit	94	89	94
Renal			
BUN	98	95	95
Chloride	98	99	99
Creatinine	71	68	79
Sodium	97	98	98
Urea	77	70	63
Hepatic			
α-1-Antitrypsin	129	154	189
Albumin	93	80	78
Alkaline phosphatase (total)	90	105	274
Bilirubin	56	56	67
Ceruloplasmin	157	189	195
Cholesterol (total)	100	127	148
Ferritin	81	32	37
Protein (total)	92	85	83
Transferrin	105	134	160
Endocrine			
1,25-Dihydroxyvitamin D		182	216
Aldosterone	90	105	274
Calcium (total)	98	95	95
Calcium (ionized)	99	98	102
Cortisol	111	238	292
Estradiol	504	3429	16494
Glucose (≥4 h fast)	98	89	91
Insulin (fasting)	72	72	134
Parathyroid hormone (intact)	52	65	54
Progesterone	337	963	1609
Thyroid-binding globulin	114	177	155
Thyroid-stimulating hormone	111	122	111
Thyroxine, free T_4	98	77	62
Thyroxine, total (T_4)	103	110	99
Triiodothyronine, total (T_3)	100	116	116
Respiratory			
Bicarbonate	85	85	85

Source: Adapted from M.L. Yarbrough, M. Stout, A.M. Gronowski, Pregnancy and its disorders, in: N. Rifai, A.R. Horvath, C.T. Wittwer (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, sixth ed., Elsevier Inc, St. Louis, MO, 2018, pp. 1655–1696; A. Gronowski, Handbook of Clinical Laboratory Testing During Pregnancy, Humana Press, Totowa, NJ, 2004; and J. Sherwin, G. Lockitch, P. Rosenthal, S. Rhone, L. Magee, E. Ashwood, et al., National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines: Maternal-Fetal Risk Assessment and Reference Values in Pregnancy, AACC Press, Washington DC, 2006, pp. 1–75 [1–3].

actions by several hormones results in an insulin resistant/glucose intolerant state characterized by decreased fasting blood glucose and increased insulin synthesis. Increased estrogen stimulates synthesis of hepatic proteins, including cortisol binding globulin, resulting in decreased cortisol clearance and the hypercortisolemic state critical for the labor process. Increased synthesis of thyroid-binding globulin results in elevated total thyroid hormone concentrations. Stimulation of thyroid-stimulating hormone (TSH) receptors by hCG results in downregulation of TSH synthesis.

Estrogen also stimulates synthesis of prolactin (which is essential for lactation), acute phase reactants such as α-1-antitrypsin and ceruloplasmin, and liver enzymes,

such as alkaline phosphatase. Increased angiotensinogen upregulates plasma renin activity and aldosterone, which promotes sodium resorption in the renal tubules. In the third trimester, increases in maternal 1,25-dihydroxyvitamin D and PTH (without corresponding increases in ionized calcium) result in enhanced intestinal absorption of calcium, critical for fetal growth. Progesterone increases during the luteal phase of the menstrual cycle and in pregnancy cause hyperventilation and mild respiratory alkalosis. To compensate for the alkalosis, plasma bicarbonate is slightly decreased; however, blood pH remains mildly alkalemic during pregnancy.

Prenatal laboratory assessment

The health of the fetus and mother is assessed before and throughout pregnancy by collecting an extensive clinical history, as well as performing a physical and prenatal laboratory testing. Currently, common laboratory testing recommended for all pregnant women includes: blood type and antibody screening, infectious disease testing, a complete blood count, immunization checks, gestational diabetes screening, urinalysis, and fetal aneuploidy and neural tube defect (NTD) screening. Women at increased risk based on clinical history are also screened for cystic fibrosis carrier status, thyroid dysfunction, and hemoglobinopathies. Table 42.2 contains a list of routine laboratory testing performed during pregnancy.

Human chorionic gonadotropin

Normal and abnormal pregnancies are detected and followed by monitoring the presence and/or changes in concentration of hCG in serum or urine.

Biochemistry of human chorionic gonadotropin

hCG is a glycoprotein hormone synthesized by the trophoblastic cells of the placenta during pregnancy. It circulates as a heterodimer of α and β subunits that are modified with both *N*- and *O*-linked oligosaccharides. hCG is part of a family of glycoprotein hormones, including LH, FSH, and TSH, characterized by common α and distinct β subunits.

In circulation, hCG has several isoforms. The main isoform in pregnancy is an intact glycosylated hCG heterodimer of ~37 kDa. In addition to intact hCG, the main isoforms found in maternal circulation include both free α and β subunits (hCGβ), nicked hCG dimer (hCGn), nicked free β (hCGβn), and hyperglycosylated hCG (hCGh). The hCGβ-core fragment (hCGβcf) is a major variant in urine. Hyperglycosylated hCG is the main isoform in circulation during early pregnancy, choriocarcinoma, and gestational trophoblastic disease (GTD). Nicked hCG is formed through enzymatic cleavage of the hCGβ between amino acids 44−45 or 47−48, and is common after term pregnancy or abortion. hCGβ is rarely seen in pregnancy, because most is rapidly cleaved and circulates as hCGβn. Some patients with germ cell tumors express hCGβ in circulation. The predominant hCG isoform in urine after 7 weeks of gestation is hCGβcf, which consists of amino acids 6−40 and 55−92 of hCGβ linked by disulfide bridges. The impact and measurement of these various isoforms are discussed below.

Function and expression of human chorionic gonadotropin in normal pregnancy

In early pregnancy, hCG maintains the corpus luteum and stimulates progesterone synthesis. Throughout pregnancy, hCG may act as a proangiogenic factor critical for implantation and maintenance of the blood supply to the fetoplacental unit. Synthesis of hCG begins shortly after implantation of the developing blastocyst into the uterus. Ultrasensitive methods can detect hCG in serum and urine up to 1 week before the predicted menstrual cycle. After implantation, hCG concentrations increase logarithmically, doubling every 1.5 days for the first 5 weeks of pregnancy. It continues to double in concentration, every 40−48 hours until it peaks between 8 and 11 weeks of gestation at ~100,000 IU/L.

Confirmation of a viable pregnancy requires serial measurements of hCG on two separate occasions 2 days apart due to significant variability in concentrations in early pregnancy. In the second and third trimesters, hCG concentrations decrease and remain steady at ~10,000 IU/L until birth. After pregnancy, hCG declines rapidly and is undetectable by 2 weeks postpartum. After an elective or spontaneous abortion in the first trimester, hCG may persist in the serum and urine up to 4 weeks due to increased concentrations and longer half-life.

Expression of human chorionic gonadotropin in abnormal pregnancy

Serial hCG measurements not only confirm normal pregnancies, but also detect abnormal ones. The minimum increase in hCG required for a viable intrauterine pregnancy (VIUP) depends on the initial value measured. Lower initial concentrations (<1500 mIU/mL) are expected to increase by at least 49% in 48 hours in a viable pregnancy, while pregnancies with initial hCG concentrations between 1500−3000 and >3000 mIU/mL are expected to minimally increase by 40% and 33%, respectively. Doubling times less than or greater than described in a VIUP are seen in nonviable and/or abnormal pregnancies. Decreased hCG doubling times are seen in

TABLE 42.2 Routine prenatal laboratory testing by trimester.

Laboratory test name	Reason for test
First trimester conception—13 weeks	
hCG	Confirm pregnancy
Complete blood count	Check for anemia and low platelet count (repeat as needed)
Genetic testing for inherited diseases	Check carrier status for certain genetic diseases
Hemoglobin A1c	Screen for type 2 diabetes (not gestational diabetes)
Hemoglobin variant analysis	Screen for a thalassemia or hemoglobinopathy[a]
Rubella titer	Screen for immunity/vaccination to rubella
HIV screening	Screen for HIV infection (all three trimesters if high risk)
Gonorrhea, chlamydia, and syphilis screening	Screen for STD infections (all three trimesters if high risk)
Hepatitis B and hepatitis C serology	Screen for hepatitis B or hepatitis C infection
Varicella foster virus testing	Screen for immunity to chickenpox
TORCH panel	Toxoplasmosis screen
Urine dip stick/screen	Screen for kidney or bladder infection, undiagnosed diabetes, or preeclampsia (at each visit)
Urine culture	Detect bacterial infection in the urinary tract
Blood typing and antibody screen	Check for blood group incompatibility between mother and fetus
Fetal aneuploidy screen	Combined screening test for trisomies 13, 18, and 21
Noninvasive prenatal testing	Fetal DNA is tested for aneuploidy and Y chromosome
Chorionic villus sampling	Confirmatory testing for positive fetal aneuploidy screening
Tuberculosis	Screen for TB[a]
TSH	Screen for thyroid dysfunction[a]
Zika virus	Screen for exposure to Zika virus[a]
Second trimester—14−26 weeks	
Glucose tolerance test	Screen for gestational diabetes
Maternal serum AFP	Screen for an open neural tube defect
Quadruple screen	Screening test for trisomies 13, 18, and 21 and ONTD
Amniocentesis	Confirmation of fetal aneuploidy by karyotyping
Third trimester—27−40 weeks	
Group B streptococcus	Screen for infection
Blood typing and antibody screen	To check for alloimmunization in Rh-negative mothers
Fetal lung maturity testing	Screen for lung maturity[a]

AFP, α-Fetoprotein; *TSH*, thyroid-stimulating hormone.
[a]*Screening for high risk patients only.*

ectopic pregnancies and miscarriages, while increased doubling time is common in molar pregnancies. Persistent low expression of hCG is seen in GTD; detection of hCGh confirms this diagnosis. Free beta hCG is expressed in malignancies like germ cell tumors, while the glycoprotein hormone alpha subunit is a tumor marker used to monitor progression of pituitary adenoma. Abnormal hCG, hCGβ, and hCGh expression is also seen with fetal aneuploidy.

Human chorionic gonadotropin immunoassays

Most commercially available hCG assays are "sandwich" immunoassays (see Chapter 12, Immunoassays). Both quantitative and qualitative assays detect some or all of the hCG isoforms in blood, urine, and body fluids to diagnose and monitor pregnancy and malignancy. Depending on the antibody epitopes, these assays may differentially recognize the distinct isoforms of hCG.

Qualitative human chorionic gonadotropin assays

Two types of qualitative assays, over-the-counter and point-of-care (POC), are used to detect hCG in urine and sometimes whole blood or serum for pregnancy testing. Most pregnancy tests are based on the principal of immunochromatography. The analytical sensitivity for these types of tests is usually 25 IU/L, a concentration typically seen at the time of missed menses. Although many devices can detect much lower concentrations (<10 IU/L), those cutoff values are less specific for pregnancy. Serum and urine hCG concentrations may be slightly elevated in men and perimenopausal and postmenopausal women due to pituitary expression leading to false-positive pregnancy testing results.

Nonpregnancy-related elevations are also seen in patients with germ cell tumors, GTDs, and heterophilic antibody interactions.

Qualitative hCG kits are also subject to false-negative results caused by multiple factors, including: hCG concentration below the limit of detection, because it was measured too early or in dilute urine (detected in samples with a low specific gravity), operator error (prematurely reading results), the hook effect (see the "Immunoassay" section for details), and/or the hCG "variant hook effect." Qualitative pregnancy testing devices differentially recognize hCG variants. Documented cases of false-negative results have been reported on numerous POC devices in women after 7 weeks of gestation, with a molar excess of hCGβcf in the urine. Some devices also demonstrated false-negative results when hCGβ, hCGβn, or hCGn were in molar excess in urine. Fewer interferences have been seen in serum specimens. Use of qualitative urine hCG testing to rule out pregnancy in emergent patients should be discouraged. While recent advances in POC kit manufacturing have reduced the incidence of the hCG variant effect, older kits are still on the market. Therefore, caution should be taken when interpreting qualitative hCG results in urine from any POC device, as they may be subject to a hook effect due to molar excess of an hCG variant like hCGβcf.

Quantitative human chorionic gonadotropin assays

Like qualitative tests, hCG quantitative assays are primarily sandwich immunoassays with antibody targets directed to different portions of the heterodimer. Most quantitative assays are performed in the clinical laboratory using serum or plasma specimens. Clinical and analytical variability exists among pregnant women and across different assays, making interpretation of results challenging. Analytical variability is due to different specificities for the intact molecule, α or β monomers, and/or other hCG variants. In addition, quantitative assays are not standardized, resulting in significant inter-method variability. While, the majority of assays are calibrated to the WHO Fourth International Standard, these are impure mixtures derived from pregnant urine and based on bioactivity instead of mass units. This mixture contains impurities as well as hCG variants like hCGn and hCGβ causing overestimation of these isoforms in some assays. Furthermore, hCG concentration in the WHO standard is based on bioactivity instead of mass units. Recommendations to correct analytical variability suggest standardization of all assays to the fifth IS, a pure material with fewer contaminating hCG isoforms. Serial hCG determinations from a patient should always be performed on the same assay. Quantitative immunoassays are also limited by interferences from heterophile antibodies or the high-dose hook effect.

Progesterone expression in pregnancy

Progesterone is produced by the corpus luteum for the first 10−12 weeks of gestation, and the placenta and adrenal glands in the remainder of pregnancy. Progesterone is critical for facilitating implantation and maintaining pregnancy and lactation. Progesterone concentrations rise steadily during early pregnancy, and deviations from this pattern indicate a poor prognosis. Serum progesterone concentrations average around 10 ng/mL in nonviable pregnancies. Progesterone concentrations of >25 ng/mL are typically associated with viable intrauterine pregnancies. Patients with ectopic pregnancies or eminent abortions often have progesterone concentrations between 5 and 15 ng/mL.

Abnormal pregnancy

Laboratory testing is critical for the identification and monitoring of abnormalities in pregnancy. The following sections will outline several abnormal conditions for mother and/or fetus and how these are assessed in the clinical laboratory.

Ectopic pregnancy

An ectopic pregnancy (EP) is any extrauterine pregnancy, with most (~97%) occurring in the fallopian tube. They account for ~2% of all pregnancies, but 6% of maternal deaths. EP is dangerous, as growth of the fetus can cause tubal rupture, maternal hemorrhage, and even maternal death. Although the exact cause of EP is unknown, risk factors include: use of assisted reproductive techniques, smoking, tubal damage from pelvic inflammatory disease, a prior sexually transmitted disease or EP, and advanced maternal age.

Clinically, it is difficult to differentiate ectopic from early intrauterine pregnancies and spontaneous abortions. Symptoms including vaginal bleeding, abdominal cramping, and pain occur in women between 5 and 10 weeks of gestation. The diagnosis is often made through a combination of sonographic and laboratory tests. Detection of an intrauterine pregnancy with 100% accuracy is not possible by transvaginal ultrasound (TVUS) until serum hCG concentrations are in the "discriminatory zone" (concentration in which a pregnancy should be seen by ultrasound) for most commercially available assays. Recent studies have challenged the use of the discriminatory zone and have suggested adopting a conservative concentration of 3500 mIU/mL. Given the variability of hCG assays and concentrations in pregnancy, each laboratory should establish its own discriminatory

hCG concentration, in which 100% of pregnancies are detected by TVUS, if a discriminatory zone is utilized in clinical practice.

Up to 31% of women with a suspected EP have an original TVUS in which the location of the pregnancy is unknown (i.e., not detected in the uterus or fallopian tube). TVUS results are correlated with serial hCG concentrations to confirm an EP. Ninety-nine percent of patients with a VIUP show an increase in hCG of at least 53% in 2 days, while 95% of patients undergoing a spontaneous abortion will show a decrease between 21% and 35% in 2 days. The majority of EPs show either hCG increases of $<50\%$ or hCG decreases of $>20\%$ in 2 days. In a minority of ectopic pregnancies, hCG expression falls outside of this expected range, making a rapid and definitive diagnosis difficult. Current research is focused on identifying biomarkers capable of diagnosing an EP at initial presentation. Candidates include markers of abnormal embryonic function (activin A and inhibin A), implantation (activin B), and/or inflammation (cytokines).

EP is treated with expectant management, chemotherapy (methotrexate injection), or surgery. The type of therapy is determined by the concentration of hCG at diagnosis, whether its concentration is increasing or decreasing, and whether there is evidence of fallopian tube rupture. After therapy, serial hCG results are followed until no longer detected in serum. In patients with high hCG and/or evidence of tubal rupture, surgery is the most effective treatment.

Molar pregnancy and gestational trophoblastic disease

There are two types of hydatidiform moles, partial and complete, that mimic pregnancy. A partial mole is often a triploid fertilization (two sperms with a viable egg) with detectable fetal material. A complete mole is typically a fertilized empty ovum with two sperms and no detectable fetal material. Moles account for ~ 1 in 1000 US pregnancies. These are usually diagnosed in the first trimester by ultrasound and/or laboratory evaluation. Partial moles are often asymptomatic and present as missed abortions, while complete moles are associated with clinical symptoms like uterine enlargement, vaginal bleeding, and less often pregnancy-induced hypertension, anemia, and hyperemesis gravidarum (HG). Ultrasound findings for a complete mole show hydropic vesicular changes in the trophoblast, giving them a grape-like appearance. Concentrations of hCG are abnormally high, and doubling times are faster than the typical 48 hours. Both types of molar pregnancies are treated with uterine evacuation and subsequent serial monitoring of serum hCG concentrations until undetectable.

About 20% of complete and 5% of partial molar pregnancies will progress to a malignant disease called GTD. Postmolar malignancies are one type of GTD, which also include noninvasive trophoblastic proliferation, invasive moles, and choriocarcinoma. GTD is most often diagnosed by detecting persistent elevation or increasing hCG in the absence of pregnancy. The predominant isoform in GTD is hCGh. Specific antibody tests for hCGh can confirm disease. GTD is treated with different chemotherapeutic courses depending upon the degree of metastasis. Hysterectomy is necessary in the most severe cases. Response to therapy is monitored by serial measurements of hCG. Persistent elevations of hCG caused by interference from "phantom hCG molecules" (i.e., trypsin cleavage products), heterophile antibodies, or pituitary hCG may occur in the absence of pregnancy and malignancy, and these should be ruled out in cases with low clinical suspicion of cancer.

Hyperemesis gravidarum

HG is a pregnancy disorder characterized by persistent nausea and vomiting (N/V), starvation (i.e., evidence of significant ketonuria), and weight loss ($\geq 5\%$ prepregnancy weight). HG is the most common cause of maternal hospital admission during the first half of pregnancy. HG is associated with extremely high concentrations of hCG, which may induce emesis. Diagnosis of HG also requires ruling out other typical causes of N/V.

Workup of HG begins with a careful clinical history, including time of onset of symptoms and severity of N/V. N/V onset after 9 weeks of gestation is likely unrelated to HG. A laboratory evaluation is helpful in cases of severe N/V. HG is commonly associated with only mildly elevated liver enzymes, bilirubin, and amylase or lipase, while these enzymes are significantly elevated in hepatitis or acute pancreatitis. Although not specific for HG, these patients may also show significant elevations in hCG and free thyroxine (FT_4) and suppressed TSH. Through its homology with TSH, hCG induces thyroid hormone synthesis and, when highly concentrated, transient hyperthyroidism.

Treatment of HG depends on abnormal laboratory results and severity of weight loss and/or dehydration. Hydration and antinausea medications are usually sufficient, although, in severe cases, hospitalization and parenteral nutrition are helpful. Transient hyperthyroidism typically resolves when hCG concentrations decrease in the second trimester and antithyroidal medications are rarely needed.

Preeclampsia

Preeclampsia is a multisystem disorder onset in the second half of pregnancy. It is defined as the presence of two

of the following: hypertension—new (bp $\geq$ 140/90 mm Hg) or worsening (an abrupt change in blood pressure or requirement for additional antihypertensive medications) in pregnancy and proteinuria ($\geq$300 mg/L urine protein/24 hours or $\geq 1+$ on a urine dipstick). In the absence of proteinuria, patients with new onset hypertension and either thrombocytopenia (platelet count $<100{,}000 \times 10^9$/L), renal insufficiency (serum creatinine $>$1.1 mg/dL), impaired liver function (liver transaminases $>$ two times the upper limit of normal), or pulmonary edema also meet the diagnostic criteria for preeclampsia. Hypertensive disorders are present in 6%–8% of all pregnancies and are a major cause of morbidity and mortality to the mother and fetus. Preeclampsia is thought to be caused by poor placentation, leading to an oxidatively stressed placenta. The clinical severity of preeclampsia varies and can present with maternal and/or fetal syndromes (intrauterine growth restriction and poor oxygenation). If left untreated, the maternal syndrome can progress to eclampsia (stroke and seizure) and/or HELLP syndrome (hemolytic anemia, liver dysfunction, and thrombocytopenia).

Preeclampsia is currently diagnosed after onset of clinical symptoms of edema, hypertension, and proteinuria after 20 weeks of gestation. Early prediction of women who will develop preeclampsia may improve maternal and fetal outcomes. However, currently there is no test that can predict preeclampsia prior to onset of clinical signs and symptoms. The most promising predictive markers include a combination of Doppler ultrasonography to monitor placental blood flow and detection of appropriate ratios of circulating proangiogenic and antiangiogenic factors, placental growth factor, and soluble fms-like tyrosine kinase-1, respectively. Despite their promise, guidelines do not recommend any screening tests for early prediction of preeclampsia other than a clinical history to identify risk factors, such as chronic hypertension, renal disease, obesity, diabetes, advanced maternal age, and personal or family history. Treatment of preeclampsia depends on severity and gestational age at onset. In late onset cases, the fetus is delivered, while, in early onset disease, a woman can be treated with antihypertensives and anticonvulsants if necessary. Fetuses that are electively delivered preterm can be treated with corticosteroids prior to delivery to facilitate lung maturity (see "Preterm birth" section) and intratracheal surfactant in the event of respiratory distress after delivery.

Thyroid dysfunction during pregnancy

Thyroid dysfunction in pregnancy can result in increased risks for the mother and fetus. Overt hypothyroidism occurs in $\sim$0.3% of pregnancies and is associated with an increased risk of miscarriage, preterm delivery, and impaired neurological development in the fetus. Subclinical hypothyroidism occurs in 2%–5% of pregnancies and may lead to cognitive deficiencies in neonates. Although controversial, most recommendations suggest screening all high risk and symptomatic pregnant women for hypothyroidism by measuring TSH. Any abnormal TSH result, considered in the context of trimester-specific reference intervals, should be followed with FT_4 testing. Some organizations recommend measuring thyroperoxidase antibodies before pregnancy or in the first trimester, and, if elevated, the mother is at risk for postpartum thyroid disease. Due to the normal decrease in TSH during pregnancy, gestational age-specific reference intervals should be used. Overt hypothyroidism in pregnancy is treated with thyroid hormone replacement therapy.

Hyperthyroidism occurs in 0.1%–0.4% of pregnancies and is associated with increased risk of spontaneous abortion, preterm delivery, preeclampsia, and thyroid anomalies in the newborn. A subnormal TSH test result should be followed with FT_4 testing. An elevated FT_4 in the presence of TSH receptor autoantibodies confirms the diagnosis of hyperthyroidism. Overt hyperthyroidism in pregnancy is treated with antithyroid drugs. In mothers with confirmed thyroid disease, fetal thyroid function can be evaluated with ultrasound to detect a goiter and AF testing for TSH, FT_4, and total T_4. AF-specific reference intervals for thyroid function tests are assay-dependent.

Hemolytic disease of the fetus/newborn

Hemolytic disease of the fetus/newborn (HDFN) is a rare immune-mediated hemolytic disorder in the fetus or neonate caused by maternal antibody attack of antigens on the surface of fetal RBCs. Its prevalence is $\sim$6/1000 live births. HDFN is caused by blood group incompatibility between mother and fetus. ABO and rhesus (Rh) factor incompatibilities are the most common form of HDFN. This is most commonly seen in the United States, when a mother's blood subtype is O while the fetus has the A blood type. Rhesus (Rh) or RhCE alloimmunization most often occurs when a woman lacking blood group antigen D or CE (Rh-negative) encounters RBCs from an RhD or CE-positive fetus. Incompatibility between mother and fetus in the expression of other RBC antigens, like C/c, E/e, A, B, Kell, Duffy, Kidd, or M, N, S, and s, can also trigger severe HDFN.

Pathophysiology

An Rh-negative and/or O blood type mother becomes sensitized to the Rh or other blood group antigens. This occurs most commonly after a blood transfusion or during a fetomaternal hemorrhage in pregnancy or at

delivery, during a spontaneous or elective abortion, or after an EP. After sensitization, antiblood group antibodies can cross the placenta in a current or subsequent pregnancy and attack antigens on the fetal RBCs, inducing hemolysis. Chronic hemolysis leads to severe fetal anemia. Compensation for this anemia involves upregulation of red cell production from the fetal bone marrow, liver, and spleen, leading to damage of these and other organs, a condition commonly referred to as erythroblastosis fetalis. Heart failure due to hypoxic injury and liver damage leading to decreased fetal albumin synthesis and reduced oncotic pressure contribute to the edematous state referred to as hydrops fetalis. If untreated, hydrops is fatal in utero or shortly after birth. After delivery, neonates with HDFN are faced with the additional challenge of removing unconjugated bilirubin without the assistance of the maternal placenta. Typically, bilirubin is conjugated and excreted through the liver; however neonates (particularly those with damaged livers or born prematurely) are not able to adequately conjugate excess bilirubin. Unconjugated bilirubin can cross the blood–brain barrier and cause encephalopathy also called kernicterus, leading to significant morbidity and, in rare cases, mortality.

Prevention

Guidelines recommend that all women should be blood typed and screened for alloantibodies (type and screen) early in pregnancy. An injection of Rh (anti-Rh) immune globulin, RhoGAM (Ortho-Clinical Diagnostics, Inc., Rochester, New York, United States) or a similar product is recommended for all Rh-negative women at 28 weeks of gestation (to prevent sensitization due to a fetomaternal hemorrhage in pregnancy) and postpartum within 72 hours if the newborn is Rh-positive. Since the implementation of routine clinical use of RhoGAM in the 1960s, the incidence of HDN has decreased significantly. Anti-Rh immune globulin should also be administered to Rh-negative women following a molar pregnancy, threatened abortion, intrauterine fetal demise in the second or third trimester, trauma to the abdomen, or a late pregnancy amniocentesis. The half-life of RhoGAM is anywhere between 16 and 30 days and varies based on dose and route, but should be administered in each subsequent pregnancy at 28 weeks.

Diagnosis and management of alloimmunization in pregnancy

Detection of significant concentrations/titers of alloantibodies during the first trimester screening test suggests that the mother is sensitized. Determination of the father's Rh type by serology and genetic testing if he is seropositive is often the next step in assessment of the fetus. The fetus will be Rh-positive if the father is homozygous for RhD and no genotyping of the fetus is necessary. A heterozygous Rh genotype in the father suggests that the fetus is at risk for developing HDFN. Fetal genotyping via PCR testing of AF or free fetal DNA testing in maternal serum is necessary. If the fetal genotype is Rh-negative, there is no risk for development of HDFN. If the fetus is RhD-positive, he or she is at increased risk for HDFN. In rare occasions, false-positive results can occur in the genetic testing of the fetus due to the presence of an RhD pseudogene. If suspected, the RhD-negative mother should be tested genetically for the pseudogene.

In a sensitized woman with an at-risk fetus, titers of alloantibodies in maternal serum are measured by the indirect antiglobulin test every 2 weeks. If the titer exceeds a previously defined cutoff, detection of fetal blood flow through the middle cerebral artery (MCA) by Doppler ultrasonography can help predict fetal anemia. Increased peak systolic blood flow through the MCA compared with a gestational age-specific median is associated with anemia of increasing severity.

In addition to ultrasonography, detection of bilirubin in AF is a sensitive predictor of the severity of fetal anemia. Liley originally determined that, in at-risk pregnancies, increased concentration of bilirubin in AF compared with normal AF of the same gestational age was directly proportional to the severity of hemolytic disease in the fetus. The concentration of bilirubin can be directly measured in AF through absorption spectrometry. The maximum absorbance of bilirubin is at 450 nm. In normal AF, the absorbance spectrum between 350 and 550 nm forms a straight line when plotted on a log–linear graph. However, in the presence of significant amounts of bilirubin, a peak is seen at 450 nm. Measurement of the linear difference between normal spectra at the absorbance of 450 nm (ΔA_{450} or ΔOD_{450} in clinical literature) is proportional to the amount of bilirubin present. The Liley chart (Fig. 42.1) plots the gestational age compared with the change in absorbance or OD at 450 nm. Three zones predicting differing severity of hemolytic disease govern management of patients. Patients whose ΔA_{450} falls within Zone I are minimally affected, while those in the upper third of Zone II or Zone III have severe disease requiring intervention. Decades later, Queenan developed a modified ΔA_{450} curve that more accurately predicts risk at an earlier gestational age, yet its use has not become mainstream. While serial measurements of AF ΔA_{450} are recommended, they are less commonly performed because of the inherent risk of pregnancy loss associated with amniocentesis. MCA Doppler ultrasonography is a noninvasive alternative. Typically, if done, serial ΔA_{450} measurements are performed every 2 weeks until delivery or intervention is necessary in antibody-sensitized women.

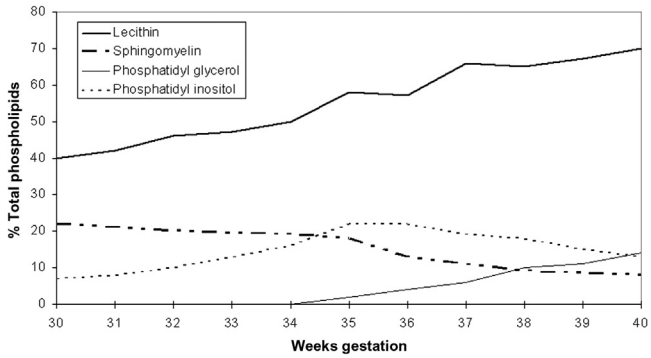


FIGURE 42.1 The Liley chart. Depiction of the relationship between bilirubin concentration as measured by ΔOD_{450} in amniotic fluid and weeks of gestation. Zone I and IIa values are associated with unaffected or mildly affected fetuses. Zone IIb or III values are associated with moderate-to-severe hemolytic disease in the fetus. *Reprinted from A.M. Gronowski (Ed.), Handbook of Clinical Laboratory Testing During Pregnancy, Humana Press, Totowa, NJ, 2004, with kind permission from Springer Science + Business Media.*

AF bilirubin measurements are affected by the presence of fetal blood. If blood is noted in AF, steps should be taken to minimize hemolysis of fetal RBCs (store in the cold and spin down cells). A Kleihauer–Betke (KB) test can be performed to check whether the blood is fetal or maternal. A KB test begins by making a slide with the bloody AF and then treating with acid. Fetal hemoglobin is resistant to acid treatment. After acid treatment, red cells in AF are stained, and fetal red cells will appear pink, while adult red cells appear as ghosts. If the blood is of fetal origin, this should be considered when interpreting results. Blood contamination of the amniotic cavity during amniocentesis takes approximately 2–3 weeks to clear from AF. Meconium contamination increases the peak at 450 nm and takes 3 weeks to clear from AF. Women carrying a fetus with severe anemia are treated with glucocorticoids to induce fetal lung maturity (FLM). These steroids falsely lower the ΔA_{450} values for up to 1 week. In the presence of all of these contaminants, delay of testing is recommended.

The gold standard for assessing fetal anemia is the measurement of hematocrit and other hematologic parameters in fetal blood. Recommendation for an in utero transfusion is based upon the fetal hematocrit, maternal clinical, and laboratory parameters along with clinical judgment. The process of ultrasound-guided fetal blood sampling (percutaneous umbilical blood sampling) is associated with a fetal death rate of 1%–2% and remains the last option when monitoring fetal anemia in a sensitized pregnancy.

Treatment of the symptomatic fetus

Treatment of fetuses with severe anemia is dependent upon gestational age and FLM. Later in pregnancy, delivery may be induced in fetuses with lung maturity and severe anemia. If lungs are immature, antenatal corticosteroids are given to stimulate maturity prior to induction or intratracheal surfactant injections given after delivery. Fetuses of <35 weeks of gestation and severe anemia may undergo an intrauterine transfusion. Perinatal survival rates are $>90\%$ for fetuses in sensitized pregnancies except for those that develop hydrops fetalis, where the survival rate is $\sim 10\%$.

Laboratory testing for fetal anomalies

Fetal congenital, structural, or chromosomal anomalies change the concentrations of certain biochemical analytes in maternal serum or AF. Powerful screening tests employing these biomarkers have been developed to estimate risk of anomalies, such as fetal aneuploidy (trisomies 13, 18, and 21), open neural tube defects (ONTDs), and fetal sex. Prenatal screening (PNS) tests are meant to identify pregnancies at risk for certain disorders that might benefit from invasive diagnostic testing. These are not diagnostic tests and should only be offered in a hospital/clinic where diagnostic testing is available. Laboratories offering PNS should communicate with the clinical care providers and patients about the risks and benefits of such testing.

Open neural tube defects

NTDs are a group of congenital anomalies involving structural defects of the brain and/or spinal cord. These effects occur early in embryonic development, typically by 6 weeks of gestation, when the neural tube fails to close. The severity and defect are dependent upon where the closure fails to occur (i.e., cranial or spinal). Cranial defects occur in the scalp and brain and are typically fatal, while spinal defects, commonly termed spina bifida, vary in severity depending upon the location of the anomaly (i.e., the higher the defect, the worse the prognosis). About 95% of spina bifida cases are open defects in which the spinal column is exposed in utero, and, at birth, 5% are closed (covered by tissue). Closed defects are usually less severe, and are rarely detected in utero by ultrasound or biochemical testing.

NTDs are caused by genetic and environmental factors. Over 90% of affected patients have no family history, suggesting a strong environmental influence. Risk factors include: geographic region, ethnicity, maternal diabetes, and diet. Of these, diet is most important. In particular, adequate intake of folic acid is essential for proper CNS development. Significant reductions in NTDs with folate intake of 400 μg/day is observed when supplementation begins prepregnancy and during the first 4 weeks. Mandatory fortification of US grain products with folate in 1998 saw the incidence of NTD drop to $\sim 1/1000$ births.

ONTDs are characterized by elevated concentrations of alpha-fetoprotein (AFP) in AF and maternal serum. AFP, a fetal protein expressed by the yolk sac and fetal liver, is part of a family of binding proteins that includes albumin. AFP concentrations are greatest in fetal serum in early pregnancy and gradually decline throughout pregnancy. The concentration of AFP in AF mirrors is of fetal serum but is 100-fold lower. In maternal serum, alpha-fetoprotein (MS-AFP) concentrations can be detected after 10 weeks of gestation and are ~4000-fold lower than in AF. MS-AFP increases throughout pregnancy and peaks at 25 weeks of gestation. Maternal serum and AF-AFP concentrations are affected by gestational age, maternal weight, early pregnancy glucose control/diabetes status, race, and the number of fetuses.

There is no clear separation in maternal serum and AF-AFP concentrations between patients with and without an ONTD; therefore it is measured as a screening test to predict women at increased risk. Screening is usually performed between 15 and 23 weeks of gestation. As noted, AFP concentrations increase with gestational age particularly during the recommended testing window. To account for this, screening test results are reported as a multiple of the median (MoM). When AFP testing is set up, laboratory-specific median values are determined for the population being tested for each gestational week.

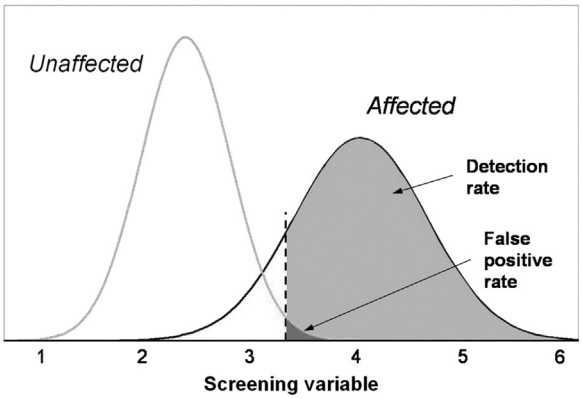


FIGURE 42.2 The effect of an arbitrary screening cutoff value (dotted line) on the detection and false-positive rates for a given disease. *Reprinted from A.M. Gronowski (Ed.), Handbook of Clinical Laboratory Testing During Pregnancy, Humana Press, Totowa, NJ, 2004, with kind permission of Springer Science + Business Media.*

The MoM calculation is as follows:

$$\frac{\text{Patient's AFP Concentration}}{\text{Median AFP Concentration for Gestational Week}}$$

The calculated MoM is adjusted to account for gestational age, maternal weight, race, diabetes status, and the number of fetuses. Prior to implementing the testing, each lab should determine the MoM cutoff for a screen positive result. The ideal cutoff optimizes detection rate while reducing false positives (Fig. 42.2). Typically, cutoffs of 2.0 and 2.5 MoM are used for both maternal serum and AF-AFP. Screening results also contain a relative risk for developing an ONTD calculated from population statistics using the patient's MoM and demographics.

Women with positive screens are referred for genetic counseling and further testing. Gestational age is confirmed by dates or ultrasound. If inaccurate, recalculation of MoM and risk is suggested. Ultrasound may also reveal other potential reasons for elevated AFP like recent fetal demise or twin pregnancy. Most ONTDs are visualized by high-resolution two-dimensional ultrasound in the second trimester. In fact, recent studies suggest that ultrasonography may have a higher detection rate than MS-AFP for NTDs. For this reason, ACOG now recommends NTD detection via high-resolution second trimester ultrasound. AF-AFP testing is often performed when ultrasound is inconclusive. All positive AF-AFP results can be checked for fetal blood contamination with an HbF immunoassay, as blood elevates AFP by 2%−3%. True-positive AF-AFP results should be followed with AF acetylcholinesterase (AChE) testing. A positive AF-AFP and AChE is 99% predictive of an ONTD.

Fetal aneuploidy

Aneuploidy, an abnormal number of chromosomes, is a common cause of miscarriage. Only a few aneuploidies are compatible with life, including trisomies 13, 18, and 21 [Down syndrome (DS)]. ACOG recommends that all pregnant women be screened for fetal aneuploidy. Screening tests identify fetuses at increased risk for DS, trisomy 18 (T18), and sometimes trisomy 13 (T13).

Down syndrome

DS is the most common chromosomal disorder among live born children, with a prevalence of ~1/800 births. DS is caused by the presence of an extra copy of a (complete or partial) chromosome 21. The disease is characterized by moderate-to-severe mental retardation, short stature, and classic facial features, including a flat occiput, upslanting eyes, small ears, congenital heart disease, and GI anomalies. The most commonly cited risk factor for DS is advanced maternal age. After 30 years, a woman's risk at term of having a child with DS is ~1 in 900, and this risk increases rapidly to 1 in 75 by age 40. Other risk factors include a having child with aneuploidy

in a previous pregnancy or detection of the genetic translocation for DS in a previous pregnancy.

In 1984 it was observed that fetal DS was associated with low MS-AFP (0.74 MoM), and this association was independent of maternal age. As with NTDs, there is significant overlap between MS-AFP concentrations in women pregnant with unaffected and DS fetuses; therefore MS-AFP was also measured to screen for DS. When combined with maternal age, MS-AFP screening detects ~50% of DS cases with a 12% false-positive rate (Table 42.3). Maternal serum concentrations of two other analytes are also altered in women with DS fetuses, unconjugated estriol (uE_3), and hCG. uE_3 is the major estrogen produced in the fetoplacental unit. It increases in maternal serum during the second and early third trimesters. Since both analytes change with gestational age, they are normalized with the MoM calculation. MS-uE_3 is lower (0.75 MoM), while hCG is higher (2.06 MoM) in women pregnant with DS fetuses. The serum triple screen (AFP, uE_3, and hCG with maternal age) is performed in the second trimester, ideally between 15 and 23 weeks of gestation, in order to assess risk for DS. Risk estimates are determined by first calculating MoMs for gestational age for each biomarker and then adjusting for interfering factors like maternal age, race, weight, and relevant clinical history. Individual laboratories determine the risk cutoff for a positive test, usually 1:270 at the time of screening. Triple screening detects 69% of DS pregnancies with a screen positive rate of 5% (Table 42.3).

Addition of a fourth analyte, dimeric inhibin A (DIA), to second trimester screening panel increased sensitivity for fetal DS. DIA is elevated in maternal serum of women with DS fetuses (≥1.77 MoM). During pregnancy, DIA is secreted by the placenta and acts to suppress FSH. Its concentrations increase slightly with gestational age; therefore it is expressed as an MoM. The quadruple screen, MS-AFP, HCG, uE_3, and DIA combined with maternal age, is the recommended second trimester test for fetal aneuploidy. Analyte MoMs combined with age and other risk factors are used to calculate risk for FDS. Adjustments in MoMs are made for maternal weight, twin pregnancies, history of prior fetal aneuploidy, and diabetes status. A lab-specific positive risk cutoff is established, which allows for detection of ~80% of FDS cases with a screen positive rate of 5%. The penta screen includes a fifth biomarker, hCGh, but currently there are limited data to support its routine use.

Gestational age should be confirmed on all positive second trimester screens. Quad screening should not be repeated, as this reduces its sensitivity. Depending on gestational age, obstetricians may elect to follow up patients with increased risk for fetal aneuploidy with noninvasive prenatal testing (NIPT)/cell-free DNA testing. AF karyotyping is the preferred confirmatory testing for all remaining screen positive patients. The benefits of diagnostic testing should be weighed against the risks of spontaneous pregnancy loss due to amniocentesis.

TABLE 42.3 Detection and screen positive rates for Down syndrome screening tests.

Screening test	Analytes	Detection rate (%)	Screen positive rate (%)
First trimester			
Maternal age		32	5
Nuchal translucency	NT	70	5
Combined test	NT, PAPP-A, and hCGβ	87	5
Second trimester			
Maternal age		32	5
Double test	AFP and hCG	60	5
Triple test	AFP, hCG, and uE_3	69	5
Quadruple test	AFP, hCG, uE_3, and DIA	81	5
Integrated (Combined test and quadruple test)	NT, PAPP-A, and hCGβ; and AFP, hCG, uE_3, and DIA	96	5
Contingent (Combined test and/or quadruple test)	NT, PAPP-A, and hCGβ; and/or AFP, hCG, uE_3, and DIA	94	5
Noninvasive prenatal testing			
Cell-free fetal DNA	DNA analysis of chromosomes 13, 18, 21, and Y	90%–99%	0.5

AFP, α-Fetoprotein; *DIA*, dimeric inhibin A; *hCG*, human C chorionic gonadotropin; *NT*, Nuchal translucency; *PAPP-A*, pregnancy-associated plasma protein-A; *uE_3*, unconjugated estriol [4].
Source: Data adapted from American College of Obstetricians Gynecologists, Practice Bulletin No. 163: screening for fetal aneuploidy, Obstet. Gynecol. 127 (2016) e123–e137.

ACOG recommends the first trimester combined screening test for all pregnant women (Table 42.3). This test combines nuchal translucency (NT) measurements taken by ultrasound and maternal serum biomarkers, hCG or hCGβ, and pregnancy-associated plasma protein-A (PAPP-A). The ultrasonographic measurement of the fluid collection at the back of the fetal neck in the first trimester, NT, is used to assess numerous fetal anomalies. DS fetuses have an increased NT for their gestational age. Alone, NT measurements detect 64%−70% of FDS with a 5% screen positive rate. Any fetus with a NT ≥ 3.0 mm is at increased risk for any aneuploidy as well as other congenital anomalies and should be offered follow-up screening or diagnostic testing. PAPP-A is synthesized by trophoblasts, and its concentration increases with gestational age. Decreased PAPP-A (≤0.43 MoM) with elevated hCG (≥1.98 MoM) and an increased NT are associated with risk of DS. The combined test offers detection rates up to 87% with 5% screen positive rates. Patients with positive screens should receive genetic counseling and diagnostic testing. The preferred confirmation (diagnostic) tests are karyotyping in the first trimester through chorionic villus sampling, performed between 9 and 11 weeks, or second trimester amniocentesis after 15 weeks of gestation. Both procedures are associated with a risk of spontaneous pregnancy loss.

Integrated testing combines both first and second trimester screening tests to assess the risk of FDS as well as ONTD. With integrated testing, the first trimester results are withheld until the second trimester testing is complete, improving DS detection to 96% with a 5% screen positive rate (Table 42.3). The improved detection rate is ideal; however disadvantages to patients are that two samples are needed, there can be significant worry during the testing period, and CVS sampling is not available for confirmation testing. Contingent screening employs a step-wise approach in which the results of the first trimester test are revealed to the patient and the patient is offered diagnostic testing if she screens positive. Intermediate and high risk screening results are followed up in the second trimester with the quad screen. The biochemical marker results are combined to generate an overall risk of fetal DS. When both trimester screening tests are used, the contingent screening approach provides a detection rate of up to 94% and a 5% false-positive rate. Disclosure of the first trimester results obviates the disadvantages of the integrated approach.

Trisomy 18

Both the first and second trimester screening approaches are capable of estimating fetal risk for T18. Screening for T18 is controversial, because the prevalence of disease and survival rate are low, ∼1/8000 live births. Both first trimester maternal serum markers (PAPP-A and hCG) are low in T18, while NT is elevated. AFP, hCG, and uE_3 are low, while inhibin is unaffected in women with T18 fetuses. The detection rate for T18 varies by study and screening test, but is ∼80% with a 0.3% screen positive rate with first trimester screening and ∼60% with a 0.5% screen positive rate when a risk cutoff of 1:100 is used for quad screening.

Trisomy 13

Also known as Patau syndrome, T13 can be detected through cell-free DNA screening. Like T18, the prevalence of T13 in live births is low at ∼1/18,000. While NT is often abnormal, neither the first of second trimester screens can sensitively detect T13.

Multifetal gestations

First and second trimester screening results are unreliable in multifetal gestations due to reduced sensitivities for detection of ONTD or fetal aneuploidy. Interpretation is particularly difficult when only one twin is affected, because maternal serum results are the average to the two fetuses. Most labs do not report risks for twin or other multiple pregnancies. Those that do report an approximate or pseudorisk calculated by dividing the patient's analyte MoM values by averages for unaffected twin pregnancies of the same gestational age.

Cell-free DNA or noninvasive prenatal testing

NIPT utilizing circulating cell-free fetal DNA is a newer screening tool for fetal aneuploidy. While there are data to support use of this testing in both high and low risk populations, it should be noted that the positive predictive value of this testing in a low risk population is decreased due to decreased disease prevalence. As with the other screening tests, decisions to terminate a pregnancy should not be made without a confirmatory diagnostic test. Small amounts of fetal nucleic acid material (3%−13% after 10 weeks of gestation) circulate in maternal serum. Several companies have patented different molecular techniques, all of which utilized next-generation sequencing of isolated cell-free fetal DNA. Along with detection of fetal aneuploidy, NIPT may also determine the sex of the infant. It is recommended that NIPT results be reported in terms of positive predictive values and posttest risks for each aneuploidy tested in the context of the percent fetal fraction isolated. Low % fetal fraction can lead to false-negative results and can occur in patients before 10 weeks of gestation, up to 23% of obese patients, and with fetal aneuploidy. Many laboratories do not report risk estimates in specimens with fetal DNA fractions of <4%, often referred to as a "no call" result. Patients with "no call" results should be offered genetic counseling,

high-resolution ultrasound, and diagnostic testing, because they are at higher risk for fetal aneuploidy. NIPT detects up to 99% of DS cases with a 0.5% screen positive rate in women whose results are reported. Like traditional screening tests, NIPT has limited utility for assessing risk of fetal aneuploidy in twin or other multiple gestations. High risk patients should be counseled on the limitations of NIPT prior to testing. In particular, despite increased sensitivity, specificity, and positive predictive values, NIPT is a screening test and does not replace traditional diagnostic techniques (i.e., karyotyping), and its results should not be used to determine treatment decisions. Furthermore, NIPT only assesses risk of fetal aneuploidy; therefore women should still be screened for an ONTD in the second trimester.

Preterm birth

Preterm birth (PTB), delivery of an infant prior to 37 weeks of gestation, is a major cause of neonatal mortality worldwide. Surviving infants suffer numerous morbidities including: respiratory distress syndrome (RDS), sepsis, neurodevelopmental delay, and behavioral problems. In the United States, approximately 10% of births were preterm in 2018, which has decreased from its peak in 2007 of 12.7%. Despite recent declines, the overall incidence has increased significantly, since the early 1980s despite advancements in prediction and prevention.

There are two types of PTB: spontaneous and induced. Some preterm deliveries are medically necessary to eliminate risk to mother or fetus. Thirty percent are spontaneous with the precipitating factor being premature rupture of membranes, while $\sim$50% of PTBs are spontaneous with intact membranes. The type of PTB depends on the underlying cause of the disease. Preterm labor is a syndrome that may be caused by one of several pathobiological processes (maternal or fetal stress, inflammation, placental abruption/decidual hemorrhage, and uterine distention) with different biochemical or epidemiological triggers. Each converges into a final common pathway involving activation of the myometrium, proteases and prostaglandins, uterine contractions and cervical dilation with or without rupture of membranes, and ultimately PTB.

Prediction of preterm birth

Prediction of spontaneous PTB is difficult, because early clinical symptoms of are typically mild and resolve without intervention. The clinical diagnosis of preterm labor, eight contractions/hour and advanced cervical dilation, often occurs too late for intervention. Epidemiologic and clinical risk factors for PTB include maternal age, socioeconomic status, lifestyle, multiple gestation, history of PTB, genetics, and race. Risk factors alone are neither sensitive nor specific predictors of those who will deliver preterm, because a majority of pregnancies resulting in PTB have no known risk factors. Cervicovaginal fluid (CVF) fetal fibronectin (FFN) is the only ACOG-approved biochemical marker for prediction of PTB in symptomatic women.

Fetal fibronectin testing

FFN is a member of the fibronectin family of matrix proteins. During pregnancy, FFN is expressed in the extracellular matrix between the maternal decidua and the fetal membranes (the choriodecidual junction) and may facilitate implantation. Cervicovaginal secretions contain FFN for the first 22 weeks of gestation until the fetal membranes fuse to the maternal decidua. By 37 weeks, FFN becomes heavily glycosylated, loses its adhesive properties, and is shed into the CVF. The presence of FFN at concentrations of $\geq$50 ng/mL in CVF of women 24−35 weeks of pregnant is associated with increased risk for PTB.

The FFN assay is FDA cleared to predict PTB in women with and without symptoms of preterm labor. In symptomatic women 24−36 weeks of gestation, FFN should be used to assess the risk of PTB within 14 days of specimen collection. Test results for FFN can also be considered in combination with other clinical information to assess the risk of PTB in asymptomatic women with singleton pregnancies of $\leq$35 weeks. Contraindications for FFN use include: advanced cervical dilation ($\geq$3 cm), rupture of amniotic membranes, cervical cerclage, moderate or gross vaginal bleeding, sexual intercourse within 24 hours, twin, or other multiple pregnancy.

The diagnostic strength of FFN lies in its negative predictive value (NPV). Symptomatic women with negative FFN results are at low risk for delivery within 7 days of testing. A positive test is less useful. The prevalence of PTB among women with symptoms of labor is $\sim$5%. The NPV of any test (even a coin flip) for a low prevalence disease is high (95% for a coin flip). Therefore the ideal predictive marker for PTB is one with a high *positive* predictive value. ACOG recommends against using FFN testing alone to determine management strategies for women with symptoms of preterm labor.

Other biochemical markers for the prediction of preterm birth

Besides FFN, the most promising marker for prediction of PTB is interleukin-6 (IL-6) in amniotic or CVF. Since the main underlying cause of PTB is infection or inflammation, cytokines like IL-6 have been implicated in the pathobiological process. The diagnostic utility of CVF IL-6 is similar to FFN (high NPV) at a fraction of the cost. New

studies are investigating panels of biomarkers associated with known genetic or etiological causes of PTB.

Prevention of preterm birth

Traditional therapies for preterm labor include hydration, bed rest, and home uterine-activity monitoring. None of these effectively prevents PTB. Drug therapy with tocolytic agents aimed at stopping contractions has limited utility. Prophylactic antibiotic therapy to prevent infectious causes of PTB showed little benefit and caused adverse effects in fetuses. Intramuscular injections of 17-hydroxyprogesterone reduce the incidence of PTB and are recommended for use in women with a prior history of PTB. While none of these reduces poor outcomes associated with PTB, they are used to prolong pregnancy long enough to treat with corticosteroids. ACOG recommends short-term use of tocolytic therapy aimed at prolonging pregnancy enough to facilitate antenatal corticosteroid treatment. One or two injections (given at least 7 days apart) of corticosteroids reduce morbidity and mortality of PTB if given at least 48 hours prior to delivery. Corticosteroids stimulate FLM.

Fetal lung maturity testing

RDS is the most common morbidity associated with prematurity. RDS occurs with incomplete fetal lung development. The incidence of RDS is increased with decreased gestational age at birth. The fetal pulmonary system is one of the last to completely develop, and risk for RDS is inversely correlated with gestational age at delivery. The final stage of fetal lung development, the alveolar stage, involves production of increasing amounts of pulmonary surfactant. In a normal airway, surfactant coats the alveolar epithelial surfaces. It decreases surface tension of the alveolar wall to prevent lung collapse during exhalation. Surfactant deficiency in RDS patients causes both lung collapse and hyperextension of alveoli, leading to fibrosis and hyaline membrane disease. The alveoli in an RDS lung are perfused, but unventilated, resulting in hypoxia, hypercapnia, and respiratory acidosis.

Pulmonary surfactant synthesis by type II pneumocytes in the fetal lung begins at $\sim$28 weeks of gestation. In type II cells, surfactant is packaged and stored in lamellar bodies (LBs). These platelet-sized vesicles made up of phospholipids (PLs; 90%) and surfactant (10%) are exocytosed to the pulmonary airway, where they coat the epithelium. During lung maturation, the PLs in LB change. Analysis of the change in PL content in AF is used to assess lung maturity.

Prevention of RDS is accomplished by preventing PTB or stimulating surfactant production with antenatal corticosteroids. Assessment of FLM status through one of several AF-based tests may assist in making clinical decisions in women with symptoms of preterm labor and/or women, whose labor is induced prior to 39 weeks of gestation. In recent guidelines, ACOG recommended against any elective induction of labor before 39 weeks of gestation; therefore FLM testing is only necessary in emergent situations (i.e., premature rupture of membranes, severe early preeclampsia requiring delivery of an infant, or HDFN). FLM tests help determine whether an NICU is needed when PTB is required. Despite this, numerous outcome studies have failed to demonstrate improvement in neonatal outcomes when FLM information is known. Furthermore, the advent of intratracheal surfactant therapy has improved outcomes for infants with immature lungs. Improved treatment options and reduced clinical utility of the testing have led to a significant decline in available FLM testing methodologies and FLM testing volumes.

Laboratory assessment of fetal lung maturity

FLM is most often assessed by testing the amount of surfactant, PL, and/or LBs in AF after 32 weeks of gestation. AF is collected through ultrasound-guided amniocentesis. Prior to testing, AF is inspected for contaminants such as blood or meconium. Contamination limits possible FLM methods. Many methods require removal of particulate matter through centrifugation. Short, slow spins are critical to avoid PL precipitation. Despite many available, highly sensitive testing options, the availability of antenatal corticosteroid therapy and intratracheal injections of surfactant has greatly reduced the number of FLM tests requested by providers.

The surfactant–albumin ratio

The surfactant–albumin (S/A) ratio increases throughout gestation proportionally with lung maturity. S/A ratio assays usually employ fluorescence polarization technology. Albumin and surfactant in AF compete for binding of a fluorescent probe. The albumin-bound probe induces a high degree of polarization, whereas the smaller surfactant bound probe has a low degree of polarization. S/A ratios are higher in mature lungs. Blood or meconium-contaminated AF may result in inaccurate results. Since the discontinuation of a commercially available S/A ratio assay, labs must develop their own, which are uncommon and not standardized.

Lamellar body count

Currently, the most common FLM method is the lamellar body count (LBC). Counting LBs in AF is a sensitive and specific alternative to S/A FLM assays. LBs pass into AF in the third trimester, and their numbers increase with FLM. They are similar in size to platelets and can be

counted on a whole blood counter. This fast and simple assay can easily be performed in any lab. LBC of $\geq 50{,}000/\mu L$ generally suggest maturity; however, cutoffs are instrument-specific. Currently, there is no standardized QC material, and LBC cannot be performed on blood or meconium-contaminated AF. Validation of LBC cutoffs is challenging, given the low incidence of RDS.

Lecithin–sphingomyelin ratio

Increases in surfactant-based PLs, like lecithin (phosphatidyl choline) and phosphatidyl glycerol (PG) in the AF compared with consistent concentrations of lipids not originating from the lung, like sphingomyelin, correlate with lung maturity. Usually, a lecithin–sphingomyelin ratio (L/S ratio) of >2.0 indicates FLM, but this is method-dependent. AF PLs are separated by thin-layer chromatography after an extraction step, and quantitated by scanning densitometry. The L/S ratio is a sensitive and specific method for assessing FLM; however, it is time-consuming, technically challenging, and nonstandardized. Availability of L/S ratio testing is uncommon, and this testing is not recommended in the presence of blood or meconium contamination.

Phosphatidyl glycerol

PG is the last surfactant-based PL to be expressed during fetal lung development (after 35 weeks), and it may help in coating the alveolar surface. Semiquantitative detection of PG in AF through an agglutination assay (AmnioSTAT-FLM-PG, Irvine Scientific) is a rapid and sensitive alternative for predicting FLM in late pregnancy. PG measurements are particularly useful in blood and meconium-contaminated AF specimens, as all other tests described above are affected by these contaminants. Because it is not usually present until late pregnancy, it has a high false-positive rate, and interpretation of the test can be inconsistent among technologists.

The appropriate management of indeterminate or immature results is unclear. Clinicians prefer to handle these individually depending on the risk to mother and/or fetus in maintaining the pregnancy. No consensus exists on when and/or whether to repeat FLM testing with an immature or intermediate result. Finally, FLM testing is not associated with improved outcomes for the neonate and has limited clinical utility.

References

[1] M.L. Yarbrough, M. Stout, A.M. Gronowski, Pregnancy and its disorders, in: N. Rifai, A.R. Horvath, C.T. Wittwer (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, sixth ed., Elsevier Inc, St. Louis, MO, 2018, pp. 1655–1696.

[2] A. Gronowski, Handbook of Clinical Laboratory Testing During Pregnancy, Humana Press, Totowa, NJ, 2004.

[3] J. Sherwin, G. Lockitch, P. Rosenthal, S. Rhone, L. Magee, E. Ashwood, et al., National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines: Maternal-Fetal Risk Assessment and Reference Values in Pregnancy, AACC Press, Washington, DC, 2006. 1-75pp.

[4] American College of Obstetricians Gynecologists, Practice Bulletin No. 163: screening for fetal aneuploidy, Obstet. Gynecol. 127 (2016) e123–e137.

Further reading

American College of Obstetricians Gynecologists, Practice Bulletin No. 97: Fetal lung maturity, Obstet. Gynecol. 112 (2008) 717–726.

American College of Obstetricians Gynecologists, Practice Bulletin No. 130: Prediction and prevention of preterm birth, Obstet. Gynecol. 120 (2012) 964–973.

American College of Obstetricians Gynecologists, Practice Bulletin No. 153: Nausea and vomiting of pregnancy, Obstet. Gynecol. 126 (2015) e12–e24.

American College of Obstetricians Gynecologists, Practice Bulletin No. 163: Screening for fetal aneuploidy, Obstet. Gynecol. 127 (2016) e123–e137.

American College of Obstetricians Gynecologists, Practice Bulletin No. 159: Management of preterm labor, Obstet. Gynecol. 127 (2016) e29–e38.

American College of Obstetricians Gynecologists, Practice Bulletin No. 187: Neural Tube Defects, Obstet. Gynecol. 130 (2017) e279–e290.

American College of Obstetricians and Gynecologists, Practice Bulletin No. 193: Tubal Ectopic Pregnancy, Obstet. Gynecol. 131 (2018) e91–e103.

American College of Obstetricians and Gynecologists, Practice Bulletin No. 202: Gestational Hypertension and Preeclampsia, Obstet. Gynecol. 133 (2019) e1–e25.

A. Gronowski, Handbook of Clinical Laboratory Testing During Pregnancy, Humana Press, Totowa, NJ, 2004.

U.H. Stenman, A. Tiitinen, H. Alfthan, L. Valmu, The classification, functions and clinical use of different isoforms of HCG, Hum. Reprod. Update 12 (2006) 769–784.

M.L. Yarbrough, M. Stout, A.M. Gronowski, Pregnancy and its disorders, in: N. Rifai, A.R. Horvath, C.T. Wittwer (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, 6th sixth ed., Elsevier Inc, St. Louis, MO, 2018, pp. 1655–1696.

Self-assessment questions

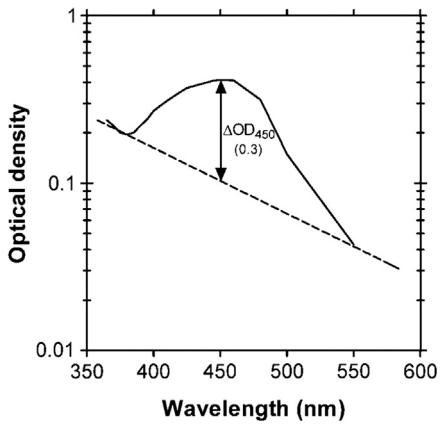
1. All of the following changes in normal laboratory values occur in pregnancy except:
   - **a.** exponential increases in progesterone, estradiol, and cortisol
   - **b.** decreases in BUN and creatinine
   - **c.** increased hemoglobin and hematocrit
   - **d.** increased liver enzymes
2. A 34-year-old female presents to the emergency room with abdominal cramping and vaginal bleeding. Her LMP was ~6 weeks ago. She has a history of an

elective abortion 12 years ago and a chlamydia infection 15 years ago. Her hCG and progesterone values over the course of a week are shown below. What is the most likely scenario?

Date	hCGβ (mIU/mL)	Progesterone (ng/mL)
10/15	493	
10/17	588	3.8
10/19	808	6.3
10/23	1136	

a. Ultrasound reveals the presence of a live intrauterine pregnancy at 6 weeks.
b. Ultrasound reveals the presence of an ectopic pregnancy in the right fallopian tube.
c. Ultrasound reveals the presence of a complete molar pregnancy.
d. Ultrasound shows no pregnancy or masses—the patient is in early menopause.

3. An RhD-negative woman is 32 weeks of pregnant with her second child and was found to have an alloantibody titer of 1:32. The profile derived from ΔOD_{450} testing of her amniotic fluid is pictured below. How should this woman be treated?



A.M. Gronowski (Ed.), Reprinted from Handbook of Clinical Laboratory Testing During Pregnancy, Humana Press, Totowa, NJ, 2004, with kind permission of Springer Science + Business Media.

a. The fetus is mildly anemic, and no intervention is required.
b. The fetus is mildly anemic, treat with antenatal corticosteroids, and induce delivery.
c. The fetus is severely anemic, treat with corticosteroids, and induce delivery.
d. The fetus is severely anemic, and perform an intrauterine blood transfusion.

4. The appropriate follow-up in a 35-year-old pregnant woman whose quadruple screen risk assessment is 1:50 for Down syndrome is:
a. No follow-up is necessary, and this is not a screen positive result.
b. Confirm gestational age with ultrasound and recalculate risk assessment.
c. Redraw maternal serum sample and repeat the quadruple screen.
d. Perform quadruple screen testing in an amniotic fluid sample.

5. Which fetal lung maturity test is most appropriate in the presence of blood contamination?
a. amniostat-FLM-PG
b. L/S Ratio
c. lamellar body counts
d. TDX-FLMII

Answers

1. c
2. b
3. d
4. b
5. a

Chapter 43

Laboratory testing in reproductive disorders

Alison Woodworth[1] and Christopher R. McCudden[2]

[1]*Department of Pathology and Laboratory Medicine, University of Kentucky Medical Center, Lexington, KY, United States,* [2]*Eastern Ontario Regional Laboratory Association, The Ottawa Hospital Department of Pathology and Laboratory Medicine, University of Ottawa, Ontario, Canada*

Learning objectives

After reviewing this chapter, the reader will be able to:

- Describe the main hormones involved in regulating the male and female reproductive systems.
- Describe how clinical laboratory testing is used to diagnose reproductive dysfunction.
- Identify the common etiologies of infertility in males and females and how they are diagnosed.

Introduction

From embryonic development to sexual maturation, the human reproductive system is controlled by hormones. The hypothalamus responds to internal and external cues to regulate pituitary secretion of stimulating hormones that target the gonads. Stimulation by pituitary gonadotrophs signals secretion of steroid hormones that provide negative feedback to the brain. The pituitary and hypothalamus also regulate steroid precursor synthesis by the adrenal glands. The hypothalamus, pituitary, and gonads act as a single unit referred to as the hypothalamic–pituitary–gonadal (HPG) axis along with the adrenal glands to regulate reproductive function.

Hypothalamic–pituitary–gonadal axis

In response to external cues and hormones, such as gonadotropins, neurotransmitters, and steroids, the hypothalamus produces gonadotropin-releasing hormone (GnRH), which signals the pituitary to release the gonadotropins, follicle-stimulating hormone (FSH), and luteinizing hormone (LH). FSH and LH are essential regulators of reproductive function in both females and males. By binding to specific receptors in the gonads, they promote conversion of steroid precursors to estradiol in females and testosterone in males. There are complex sex-specific differences in the type, quantity, and timing of steroid production in response to gonadotropins. It should be noted from the outset that hypothalamic hormones (e.g., GnRH) occur at very low concentrations with very short half-lives, and are thus not readily measurable in peripheral circulation. Consequently, pituitary hormones, such as LH and FSH, are used as surrogates for hypothalamic GnRH in the workup of most reproductive disorders.

Male reproductive endocrinology

The male endocrine sex organs are the testes, which contain two hormone-producing cell types, Leydig and Sertoli. Leydig cells synthesize the male sex steroids (androgens). Sertoli cells make sperm and produce hormones called inhibins that provide negative feedback to the brain. Sertoli cells also establish and maintain the "blood–testes barrier," facilitating a unique luminal fluid composition as compared with blood. The testes rely on signaling and regulation by pituitary hormones, such as GnRH.

Control of testicular function

GnRH is a decapeptide secreted in pulsatile fashion with peaks occurring approximately every 90–120 minutes. Pulses of GnRH are critical for maintaining normal concentrations of gonadotropins. The specificity for release of each is a function of the amplitude and frequency of GnRH signaling. Low-frequency pulses stimulate FSH release, while high-frequency pulses promote LH release. As a consequence, concentrations of GnRH, FSH, and LH are variable at any given time. This secretion pattern is often diagnostic, as pulsatility may be lost in pathophysiological states. Accordingly, interpretation of blood FSH and LH measurements requires an understanding of the

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00043-0

underlying physiological conditions. Structurally, both FSH and LH are secreted as heterodimers consisting of α- and β-subunits with homology to TSH and hCG. Consistent with other pituitary glycoprotein hormones, the α-subunit is the same, while the β-subunit differs and confers the biological specificity.

Once released into circulation, LH binds G-protein-coupled receptors (GPCRs) on the Leydig cell surface, stimulating androgen synthesis [principally testosterone, but also dihydrotestosterone (DHT)]. The increase in circulating testosterone inhibits secretion of GnRH from the hypothalamus and LH from the pituitary (Fig. 43.1). FSH binds to a related GPCR on Sertoli cells and promotes spermatogenesis. Spermatogenesis requires high concentrations of testosterone in the seminiferous tubules, and is maintained by proximity to the Leydig cells and androgen-binding protein from the Sertoli cells (steroids are able to cross the blood–testes barrier). FSH also stimulates Sertoli cells to secrete inhibin. The two isoforms of inhibin (A and B) are heterodimeric in structure and composed of dissimilar subunits. Inhibin B is the predominant isoform produced in males although both inhibins signal the pituitary to downregulate FSH secretion.

Androgens

In males, the testes and adrenal glands secrete several different androgens including testosterone, dihydrotestosterone, androstenediol, dehydroepiandrosterone (DHEA), and dehydroepiandrosterone sulfate (DHEA-S). Like all steroid hormones, androgens are synthesized from cholesterol through a series of compartmentalized enzymatic conversions (Fig. 43.1—male steroidogenesis). The most potent androgen, DHT, is produced largely through peripheral conversion of testosterone. DHT has limited diagnostic utility, because circulating concentrations do not reflect intracellular levels. Functionally, DHT is an important steroid in the development of benign prostatic hyperplasia (BPH). BPH is treated with 5α-reductase inhibitors, which block conversion of DHT from testosterone. Androstenediol and other testosterone precursors have limited bioactivity due to weak affinity for androgen receptors. In clinical and diagnostic terms, testosterone is the most important androgen; in cases of anabolic steroid abuse, other androgens may be important.

Testosterone

Testosterone stimulates differentiation of the male fetal genital tract, establishes and maintains male secondary sex characteristics, and promotes spermatogenesis. The testes secrete ~4–10 mg of testosterone daily, accounting for 95% of total circulating concentration; the remaining 5% is produced by peripheral conversion of the adrenal androgens, such as DHEA. Testosterone exhibits diurnal variation in adult men with the highest concentrations occurring in the morning, followed by a gradual decline throughout the day. Testosterone levels reach their

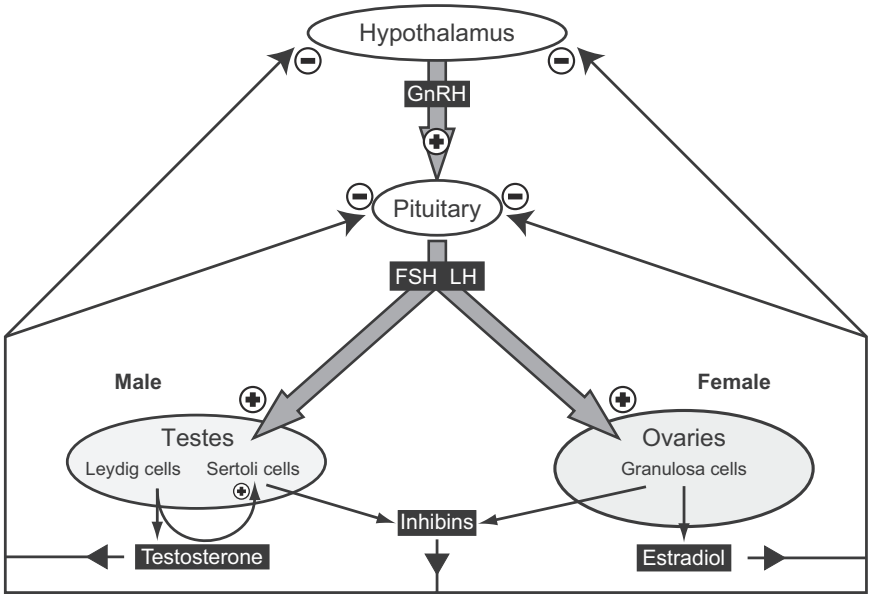


FIGURE 43.1 Hypothalamic–pituitary–gonadal axis. The hypothalamus secretes gonadotropin-releasing hormone in response to environmental and internal cues stimulating gonadotropin (follicle-stimulating hormone and luteinizing hormone) release from the pituitary. In males, luteinizing hormone stimulates testosterone production by Leydig cells. Follicle-stimulating hormone in combination with testosterone promotes spermatogenesis within the Sertoli cell-enclosed seminiferous tubules. In females, follicle-stimulating hormone and luteinizing hormone regulate ovarian production of estradiol in the menstrual cycle. In both males and females, inhibins and sex steroids both provide negative feedback to the hypothalamic–pituitary axis.

nadir in the evening and during the first few hours of sleep. This rhythm is less pronounced with advancing age and in men with testicular failure.

Although a small amount (~2%) of free testosterone is found in circulation, most is bound to carrier proteins including albumin and sex hormone-binding globulin (SHBG). SHBG is synthesized by the liver and binds ~45%–60% of circulating testosterone with high affinity (Kd ~1 nM). The remaining portion of testosterone (35%–55%) is transported bound to albumin with weak affinity (Kd > 1000 nM). SHBG synthesis is increased by estrogen or excess thyroid hormones, and decreased by exogenous androgens, glucocorticoids, growth hormone, and hypothyroidism. Similarly, liver damage may affect synthesis of SHBG and accordingly measurement of total testosterone. Although the free fraction of any hormone is considered to be the biologically active form, there is considerable evidence that the albumin-bound and free portions of plasma testosterone, termed "bioavailable testosterone," represent the biologically active fraction. Provided the affinities above, consider that circulating concentrations of testosterone (Table 43.1) will typically be below the affinity of albumin, such that it will be biologically available.

Measurement of testosterone

There are several different methods available to measure total testosterone including equilibrium dialysis, direct immunoassay (chemiluminescent or radioimmunoassay), and chromatography/mass spectrometry (LC–MS/MS and GC–MS). Isotope dilution mass spectrometry is the reference method for testosterone measurement, although most labs utilize immunoassays. Immunoassays rely on chemical blocking agents that displace testosterone from SHBG and albumin, followed by competitive binding with labeled hormone. There are a number of challenges surrounding measurement of testosterone: (1) assay specificity (cross-reactivity with related steroid hormones); (2) analytical sensitivity (detection of low concentrations of testosterone); and (3) appropriate reference intervals and partitions for men, women, and children. In particular, total testosterone immunoassays are subject to imprecision at low concentrations and when SHBG concentrations are abnormal. Thus detection of androgens in women, children, and some aging men requires more sensitive methods (e.g., LC–MS/MS or second generation immunoassays) to accurately measure testosterone.

Free and bioavailable testosterone

In women and hypogonadal males, second line testing relies on measurement or estimation of free/bioavailable testosterone; these are not widely used in children. The gold standard for free testosterone measurement is equilibrium dialysis/high-performance liquid chromatography–tandem mass spectrometry (ED/LC–MS/MS), although ultrafiltration is a popular alternative used to

TABLE 43.1 Free and bioavailable testosterone measurement limitations.

Method	Limitation
Equilibrium dialysis/high-performance liquid chromatography–tandem mass spectrometry	Expensive and limited availability Results vary with pH, temperature, and dilution Laborious (dialysis alone can take >12 h)
Testosterone analog immunoassays	10-fold negative bias against reference method Not recommended
Free androgen index	Inaccurate due to assumptions about linear relationship between parameters in calculation Not recommended
Calculated free testosterone and bioavailable testosterone	Requires accurate methods unpinning calculation (subject to variation in testosterone and SHBG) Several physicochemical assumptions about dissociation constants (binding affinity, albumin concentrations) Inconsistent reference intervals Many different available equations requiring local validation >50% difference in values reported on proficiency surveys
Ammonium sulfate precipitation	Labor intensive Results vary with temperature Interindividual variation (between staff running assay)

SHBG, Sex hormone-binding globulin.

reduce the amount of time needed for analysis. Direct radioimmunoassay (RIA) testosterone analog methods that rely on competitive binding of a nontestosterone analog to an immobilized antibody have limited clinical utility (Table 43.1). There are also protein precipitation methods, where ammonium sulfate is used to crash protein out of solution leaving behind, in theory, only bioavailable testosterone.

As an alternative to direct measurement, free or "bioavailable" testosterone may also be calculated from total testosterone and SHBG based upon testosterone constant for binding to SHBG using one of several formulae (equations include those from Sodergard, Vermeulen, Nanjee, Sartorius, and Ly). Each of these has limitations in assumptions about binding and kinetics of testosterone and binding proteins (Table 43.1). Of historical note, free androgen index (FAI) calculation (% ratio of testosterone/SHBG) was once used, but is considered inaccurate, because it does not account for the nonlinear relationship between total testosterone and SHBG. Most laboratories have discontinued reporting the FAI.

While the free/bioavailable testosterone methods have their challenges, there have been substantial efforts toward standardization of total testosterone. Currently, the Centers for Disease Control and Prevention (CDC) runs a standardization program, whereby a number of labs with LC−MS/MS methods are certified, provided they achieve results within ∼6% of their reference method. CDC also provides target values for the College of American Pathologists proficiency surveys, which has improved overall agreement of both LC−MS/MS and immunoassays. From a clinical standpoint, practice guidelines recommend measurement of serum total testosterone in a morning collection to screen for deficiency. Bioavailable and/or free testosterone are then used as confirmatory tests.

Dehydroepiandrosterone and dehydroepiandrosterone sulfate

Most DHEA (∼4 mg/day) and DHEA-S (∼10 mg/day) are synthesized by the adrenal glands. Production of these androgens increases just before onset of puberty, peaks in the third decade, and then gradually declines. DHEA-S, a sulfated derivative accounting for >99% of circulating DHEA, is formed largely in the adrenal cortex, but also peripherally through the actions of liver and kidney sulfotransferases.

Both DHEA and DHEA-S are transported primarily bound to albumin. While they are partially influenced by ACTH and the immune system, the regulation of DHEA and DHEA-S is not well understood. Combined DHEA and DHEA-S are the most abundant steroids in the body, where they bind weakly to the androgen receptor and serve as precursors for other steroids. In females, DHEA and DHEA-S contribute approximately half of the total circulating testosterone either directly or through conversion to other testosterone precursors. DHEA-S is preferred for laboratory analysis because of its relative abundance, little to no diurnal variation, and longer half-life (7−10 hours and 15−30 minutes for DHEA-S and DHEA, respectively). DHEA-S measurements are useful when investigating precocious puberty in males, and hirsutism, amenorrhea, or infertility in women.

Male reproductive development

In males, numerous genes on the Y chromosome promote development of indifferent gonads into testes. Embryonic testes control male sexual differentiation through synthesis of various protein and steroid hormones. Testosterone stimulates the Wolffian ducts to develop into the vas deferens, epididymis, and seminal vesicles, and for the differentiation of external male genitalia. After birth, concentrations of the gonadotropins and sex steroids are initially elevated, but then decrease during the first few days of life. Over the next several weeks, plasma concentrations of LH and FSH are higher than during the rest of childhood in both sexes. In males, testosterone is also higher in infants than in older children, suggesting that gonadotropins stimulate its synthesis. Sex steroids, LH, and FSH peak at 2−3 months of age and then decrease to low concentrations that persist throughout childhood. Before puberty, the hypothalamus and pituitary are extremely sensitive to negative feedback from sex steroids and inhibin. Before maturation of the HPG axis, adrenal synthesis of androgens begins to increase. This increase in adrenal androgens (andrenarche) is responsible for the onset of pubic and axillary hair growth.

The increased and intermittent release of GnRH from the hypothalamus causes surges in LH and FSH. The diurnal variation gradually wanes, and the pulsatile release of the gonadotropins occurs at all times during the day. At puberty, the earliest visible sign of increasing LH and FSH secretion is an increase in testicular size. At this time, the hypothalamus and pituitary are less sensitive to the negative feedback allowing the concentrations of gonadotropins and sex steroids to rise steadily, leading to the development of secondary sex characteristics.

The decline of testosterone with advancing age is often referred to as "andropause." Unlike the female menopause where there is a complete absence of estrogen with associated clinical symptoms, andropause involves only a slight decrease in testosterone with less obvious clinical findings. These consist of declines in sexual function, bone mineral density, muscle mass, strength, and cognitive function. Concentrations of SHBG increase with age, resulting in an increase in total testosterone, decreased bioactive testosterone, and increased FSH and LH.

Male reproductive disorders

Reproductive disorders in the male can be divided into five categories including hypergonadotropic (primary) and hypogonadotropic (secondary) hypogonadism, defects in androgen action, impotence, and gynecomastia (Table 43.2). In primary hypogonadism, testosterone is low, and gonadotropins (FSH and LH) are elevated. In contrast, secondary hypogonadism results from low gonadotropins.

Hypogonadism

Male hypogonadism is a clinical diagnosis defined by the inability of the testes to produce testosterone with or without the presence of a low sperm count. Hypogonadism results from disruption of the hypothalamic−pituitary−testicular axis, and diagnosis requires persistent measurably low testosterone output. Clinical features of hypogonadism vary depending on the extent of the disorder. Decreased spermatogenesis is associated with infertility and decreased testicular size. Decreased testosterone synthesis manifests in a variety of symptoms depending on time of onset. A complete deficiency of testosterone in early development results in impaired sexual differentiation in the fetus, where a genotypic male will possess female external genitalia. A partial testosterone deficiency causes incomplete virilization. Males will not develop secondary sexual characteristics at adolescence if testosterone is decreased before puberty. In adults, testosterone deficiency causes decreased libido, muscle mass, and bone mineral density (similar to andropause). Low plasma total testosterone results support the diagnosis of male hypogonadism (Table 43.2). The American Society for Reproductive Medicine (ASRM) Practice Committee guidelines recommend confirmation of low results by repeating morning measurement and/or using free/bioavailable testosterone in cases where SHBG concentrations are abnormal or results are equivocal. The Endocrine Society recommends the testosterone measurement method to be certified by an accuracy-based standard, such as the CDC Hormone Standardization Program for Testosterone. GnRH stimulation testing may also be performed (see hypergonadotropic hypogonadism below).

In hypergonadotropic hypogonadism (primary hypogonadism), there is insufficient steroid production by the gonads. Lack of negative feedback from testosterone leads to increased secretion of FSH and LH by the pituitary. Causes of primary hypogonadism include Klinefelter syndrome (47, XXY karyotype), testicular damage from chemotherapy, radiation, disease, drugs, trauma, and aging. Significant increases in gonadotropins indicate primary hypogonadism even when testosterone is only slightly decreased.

Hypogonadotropic hypogonadism results from low or absent secretion of FSH and LH from the pituitary (secondary hypogonadism) or impaired secretion of GnRH from the hypothalamus (tertiary hypogonadism). Causes include tumors that interfere with hypothalamic−pituitary axis function, panhypopituitarism, hyperprolactinemia, GH deficiency, hypothyroidism, craniopharyngioma, malnutrition, and anorexia. Recently developed immune checkpoint therapies (anti-PD-1/PD-L1 and anti-CTLA-4) are a newer cause of HPA disruption, which can cause secondary hypogonadism. Hypogonadotropic hypogonadism is identified by measuring decreased circulating testosterone in combination with normal or reduced LH and

TABLE 43.2 Diagnosis of androgen excess disorders.

Feature	Comment
Detailed clinical history	Timeline of symptoms
PCOS	Androgen excess
	Oligo or amenorrhea
	Imaging to establish polycystic ovary morphology
Exclusion criteria	Thyroid abnormalities
	Hyperprolactinemia
	Atypical CAH (non-21-OHase deficiency)
Patient-specific exclusion criteria	Acromegaly
	Cushing's syndrome
	Neoplastic secretion of androgens

CAH, Congenital adrenal hyperplasia; *PCOS*, polycystic ovarian syndrome.

FSH concentrations. GnRH stimulation testing may also be performed to determine if the pituitary is responsive. Where LH and FSH increase with GnRH administration, hypergonadotropic hypogonadism may be ruled out. Therapy depends on identifying the underlying condition.

Defects in androgen action

In addition to primary and secondary causes, male reproductive disorders can also result from defective androgen action. These are divided into two categories: (1) defects in the testosterone conversion pathway; and (2) androgen receptor defects.

For testosterone to exert its full biological effect, it must be enzymatically converted to DHT by 5α-reductase. Congenital deficiency of 5α-reductase leads to male pseudohermaphroditism, ambiguous genitalia, and a microphallus. At puberty, normal concentrations of testosterone allow the development of non-DHT-dependent secondary sex characteristics (increased muscle mass, laryngeal enlargement, and enlargement of phallus). In these cases, male children may be raised as females until puberty, necessitating early diagnosis to avoid gender confusion. These individuals are also at risk for development of testicular cancer. Deficiency of 5α-reductase is verified by measuring an increased testosterone:DHT ratio in affected individuals.

Androgen resistance, also referred to as testicular feminization, is the most severe defect in androgen action, and may be caused by any of 150 different mutations in the androgen-receptor gene. These X-linked recessive receptor defects range from partial androgen response to complete androgen insensitivity. Affected individuals have a 46, XY karyotype and bilateral testes, with female external genitalia (a vagina that ends in a blind pouch). The breasts undergo normal (female) development at puberty. A diagnosis of androgen resistance is suggested by normal concentrations of testosterone for a male in a phenotypic female.

Erectile dysfunction

Erectile dysfunction (ED), the persistent inability to develop or maintain an erection, has many causes. Endocrine-related pathologies of ED include diabetes mellitus, hypogonadism, hyperprolactinemia, adrenal insufficiency, feminizing tumors, and both hypothyroidism and hyperthyroidism. Nonendocrine causes of ED include psychological factors, autonomic neuropathy, vascular deficiency, or drug-related side effects. Clinical history is the most effective means to assess the underlying cause of ED and may prompt laboratory evaluation; tests include testosterone, prolactin, and thyroid hormones. As ED is relatively common in men with diabetes mellitus, fasting plasma glucose should also be performed.

Gynecomastia

Benign proliferation of glandular, not adipose, tissue of the male breast, gynecomastia, can occur any time from infancy to old age. It is caused by an imbalance between estrogen and androgen metabolism and results in an increased estrogen:androgen ratio. In neonates, gynecomastia is a side effect of the normally high estrogen concentrations found in pregnancy. Once exposure to estrogen stops at birth, the condition resolves spontaneously. During puberty, gynecomastia occurs when high concentrations of pituitary gonadotropins favor estrogen precursor secretion over testosterone. This condition usually resolves as puberty ends, and steroid synthesis of androgens predominates. In men over 50, gynecomastia is associated with increased aromatase (the enzyme that converts testosterone to estradiol) present in adipose tissue and decreased testosterone synthesis by aging testes.

Gynecomastia can also be induced by several drugs. Prolactin stimulators, such as phenothiazines, methyldopa, and reserpine, suppress gonadotropins and testosterone, whereas spironolactone and anabolic steroids increase aromatization of testosterone to estradiol. Digoxin and phytoestrogens from marijuana act as ligands for the estrogen receptor, such that there is effectively an increased estrogen:testosterone ratio. After excluding pubertal or drug-induced causes, laboratory evaluation for gynecomastia includes measurement of plasma testosterone, estradiol, LH, prolactin, and hCG. In hypogonadic gynecomastia, testosterone is low, and LH and estradiol are elevated. Measurement of hCG is used to detect germ cell tumors that may stimulate estrogen synthesis.

Female reproductive endocrinology

The primary female endocrine sex organs are the ovaries. Ovarian follicles contain oocytes, which are surrounded by several layers of steroid producing cells, the inner granulosa cell layer and the outer theca layers. The theca layers produce steroid precursors, and the granulosa cells convert these precursors to estrogen. Progesterone is produced from the corpus luteum, which develops from ruptured follicles after ovulation. The ovaries function in combination with the hypothalamus and pituitary to regulate the female reproductive system (Fig. 43.1).

Control of ovarian function

As in males, the hypothalamus and pituitary control the primary sex organs in females. Hypothalamic GnRH stimulates the anterior pituitary to release FSH and LH, which promote follicle development and ovulation, respectively; LH also stimulates synthesis of progesterone by the corpus luteum. Ovarian steroids normally inhibit

the secretion of FSH and LH, but, under certain circumstances, exhibit positive feedback, resulting in a release of the pituitary gonadotropins. Ovaries also produce inhibins (A and B), which suppress secretion of pituitary FSH; inhibins may be measured in the context of ovarian tumors and determining ovarian reserve (see below).

Ovarian steroids

The ovaries synthesize sex steroids in a series of enzyme-catalyzed steps from cholesterol (Fig. 43.2). In females, the ovaries are the primary source of estrogens, progestins, and androgens, although the adrenal glands can also produce many of these steroids.

Estrogens

Estrogens are the primary female sex hormones responsible for the development of the secondary sex characteristics, regulation of the menstrual cycle (in conjunction with progesterone), breast and uterine growth, and the maintenance of pregnancy. Although more than 20 different estrogens have been identified, only three have known clinical relevance: estrone (E1), estradiol (E2), and estriol (E3). Of these, estradiol is the predominant estrogen produced by the ovaries, making it a good marker of ovarian function. Estrogen is also produced in small amounts by the adrenal glands and peripheral conversion of androgens by aromatization to estrone. During pregnancy, estriol,

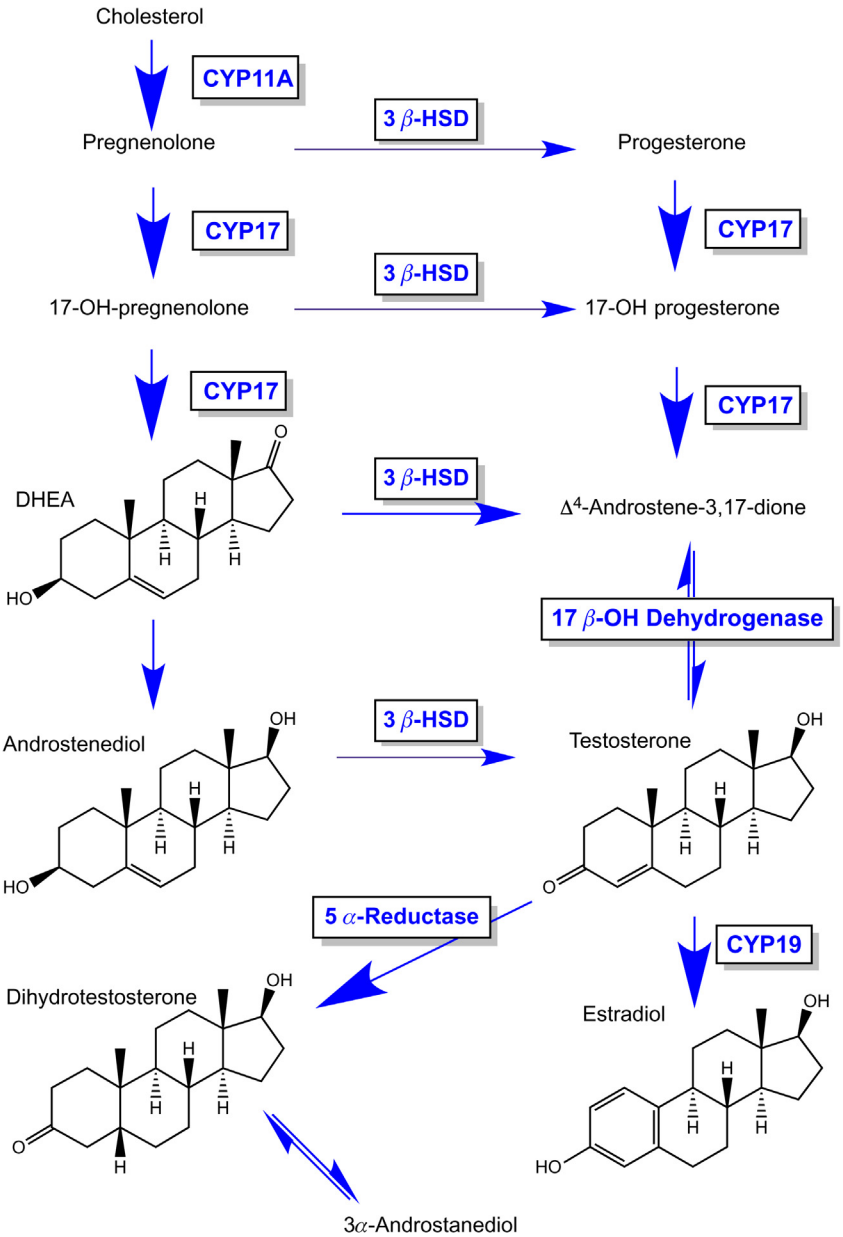


FIGURE 43.2 Overview of androgen synthesis pathway with important converting enzymes.

synthesized in the placenta from fetal 16α-OH-DHEA-S, is present in high concentrations in maternal serum. In nonpregnant, premenopausal women, the concentration of estrogen varies in a temporal pattern throughout the menstrual cycle (Fig. 43.2).

Estrogen exerts its effects by binding to nuclear estrogen receptors. The receptor−ligand complex binds specific response elements in target genes and induces transcription. In addition to nuclear receptors, there are also membrane-bound G-protein-coupled estrogen receptors (mERs). These receptors act nongenomically through complex and incompletely understood mechanisms. Circulating estrogen is metabolized by the liver to estrone, and subsequently to catechol estrogens (2-hydroxyestrone, 2-hydroxyestradiol, and 2-hydroxyestriol) or estriol. Estrogen metabolites are glucuronidated and sulfated rendering them water soluble facilitating excretion in bile or urine.

The majority (97%) of estradiol is bound weakly to albumin (~60%) and tightly to SHBG (~40%) in circulation. Estrogen stimulates the synthesis of binding globulins; thus SHBG concentrations are nearly twice as high in females as in males. Both free and albumin-bound estradiol are considered biologically active, although the clinical significance of this is unclear.

Measurement of estrogen

Circulating estrogens are most commonly measured by immunoassays. As with testosterone, estradiol immunoassays are effective for the majority of cases (infertility assessment in postpubertal females). However, estradiol immunoassays remain insensitive at low concentrations and are ineffective at assessing abnormalities in children and males. RIAs are capable of detecting lower estradiol concentrations, but are slow, labor intensive, and use radioactive reagents. Both RIAs and rapid immunoassays yield variable results between different manufacturers, and measurement methods must accommodate a wide dynamic range (1−3000 pg/mL).

LC−MS/MS methods to measure estrogens have improved low-end sensitivity and are faster than RIAs, but are technically demanding. LC−MS/MS measurement of estradiol is sensitive to interferences, requires extraction, may require derivatization, and is not readily implemented without substantial expertise. The Endocrine Society has called for standardization and improved sensitivity for estradiol assays. Based on the success of testosterone standardization (see "Measurement of testosterone" section), the CDC has defined an HPLC−MS/MS reference method, standard calibrator, and agreement on a certified reference material (SRM-971) for estradiol. A congressionally backed consortium (Partnership for Accurate Testing of Hormones) that includes the American Association for Clinical Chemistry has the goal of standardization and harmonization of estradiol (among other hormone tests).

Progesterone

Progesterone acts in concert with estrogen to regulate the menstrual cycle and prepares the uterine endometrium for implantation of a fertilized egg. Progesterone also regulates glandular development of the breast and maintains the uterus during pregnancy. The adrenal glands, ovaries, and placenta all synthesize progesterone. The circulating concentration of progesterone in nonpregnant, premenopausal women varies throughout the menstrual cycle (Fig. 43.3).

In the corpus luteum, LH and FSH control secretion of progesterone. Once in circulation, progesterone is transported bound to corticosteroid-binding globulin and albumin, with approximately 2%−10% remaining free. Similar to other steroids, progesterone exerts its effects by binding to a specific nuclear receptor in target cells where it modifies gene transcription. Progesterone has a rapid clearance rate and is metabolized in the liver by conversion into one of three classes of compounds: pregnanediones, pregnanolones, and pregnanediols. Conjugation of these metabolites with glucuronate facilitates their excretion by the kidney.

Female reproductive development

In genetic (XX) females, the absence of a Y chromosome allows the gonads to differentiate into ovaries. The absence of testosterone and anti-Müllerian hormone causes the Wolffian ducts to degenerate and the Müllerian ducts to differentiate into the female reproductive tract. At mid-gestation, the developing ovaries contain approximately 7 million oocytes arrested at the first meiotic prophase. An interesting phenomenon related to this early oocyte development is transgenerational epigenetic inheritance, where epigenetic changes (e.g., DNA methylation) in the mother are potentially transmitted to her grandchildren. It is hypothesized that this occurs by differential DNA methylation of the oocytes developing within the female fetus. Oocytes remain in the first meiotic phase until just before ovulation. The number of oocytes declines steadily over a woman's lifetime, and, by menopause, there are few remaining.

As in the males, gonadotropin concentrations in female children are low. These low concentrations of FSH are sufficient to stimulate the granulosa cells to express aromatase, which is required for estrogen synthesis and the follicles to secrete inhibin. LH appears to have little function in female children prior to puberty. Endocrine changes at puberty are initiated by an increase in the frequency and amplitude of the gonadotropins and a corresponding increase in the concentration of ovarian sex hormones. Increased estrogen, the development of the secondary sex characteristics, and onset of menses

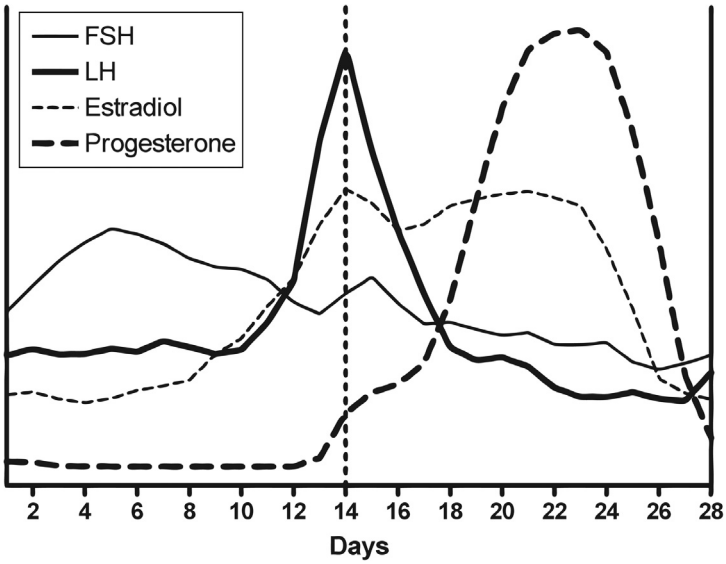


FIGURE 43.3 Relative changes of hormone concentrations throughout the menstrual cycle. Days 1–14: follicular phase. Days 15–28: luteal phase. Ovulation is indicated at Day 14.

(menarche) mark the defining event of puberty in females. The age of puberty onset in females is largely determined through a combination of genes, though other factors contribute. For example, a certain amount of adipose tissue is required to permit puberty to occur (leptin is needed for puberty to occur, but by itself will not induce puberty).

Menstrual cycle

The menstrual cycle is a series of tightly controlled events by the female HPG axis. Its median length is 28 days and consists of two phases (follicular and luteal) divided by ovulation at ~14 days. The hormonal changes observed during the menstrual cycle are shown in Fig. 43.3.

Follicular phase

During the follicular phase, concentrations of ovarian steroids are low, permitting an increase in both FSH and LH. The rapid rise of FSH stimulates maturation of a single follicle. As the follicle develops, synthesis of estradiol increases to stimulate proliferation of the endometrial lining of the uterus. Increasing estradiol concentrations inhibit FSH release by the pituitary, and thus its concentrations decline in the late follicular phase.

A day prior to ovulation, circulating estradiol concentrations peak. At this point, there is a switch from negative to positive feedback control by estradiol, resulting in an LH surge (Fig. 43.3). The LH surge indicates that ovulation will occur within 36 hours and marks the transition from the follicular to the luteal phase of the menstrual cycle. The LH surge can be detected in the urine 24–36 hours prior to ovulation, but does not confirm that ovulation actually occurred.

Luteal phase

In the luteal phase, the ruptured follicle transforms into a corpus luteum. In this process, granulosa cells luteinize (enlarge and appear yellow), as they accumulate cholesterol needed to synthesize progesterone. Negative inhibition by ovarian hormones induces a gradual decrease in FSH and LH. Progesterone and estradiol concentrations peak ~8 days after ovulation. In the absence of fertilization, progesterone concentrations begin declining around day 23 of the cycle, and the endometrium sloughs off as menstruation ensues (Fig. 43.3).

Female reproductive disorders

There are myriad female reproductive disorders, all of which can result in the persistent inability to ovulate. These may manifest as disturbances in the menstrual cycle, infertility, or syndromes of androgen excess. Female reproductive disorders can be broadly classified as hypothalamic–pituitary dysfunction, congenital disorders, and androgen excess (Table 43.3).

Hypothalamic–pituitary dysfunction

Abnormal secretion of hypothalamic or pituitary hormones can adversely affect the female reproductive system. Drugs, physical or psychological stress, tumors, autoimmune disease, and trauma all disrupt the female HPG axis. Diagnosis of hypothalamic–pituitary causes of reproductive dysfunction requires a detailed clinical history and may include imaging studies to rule out neoplasms. Patient history identifies onset of symptoms, comorbidities, and psychosocial and drug-related causes.

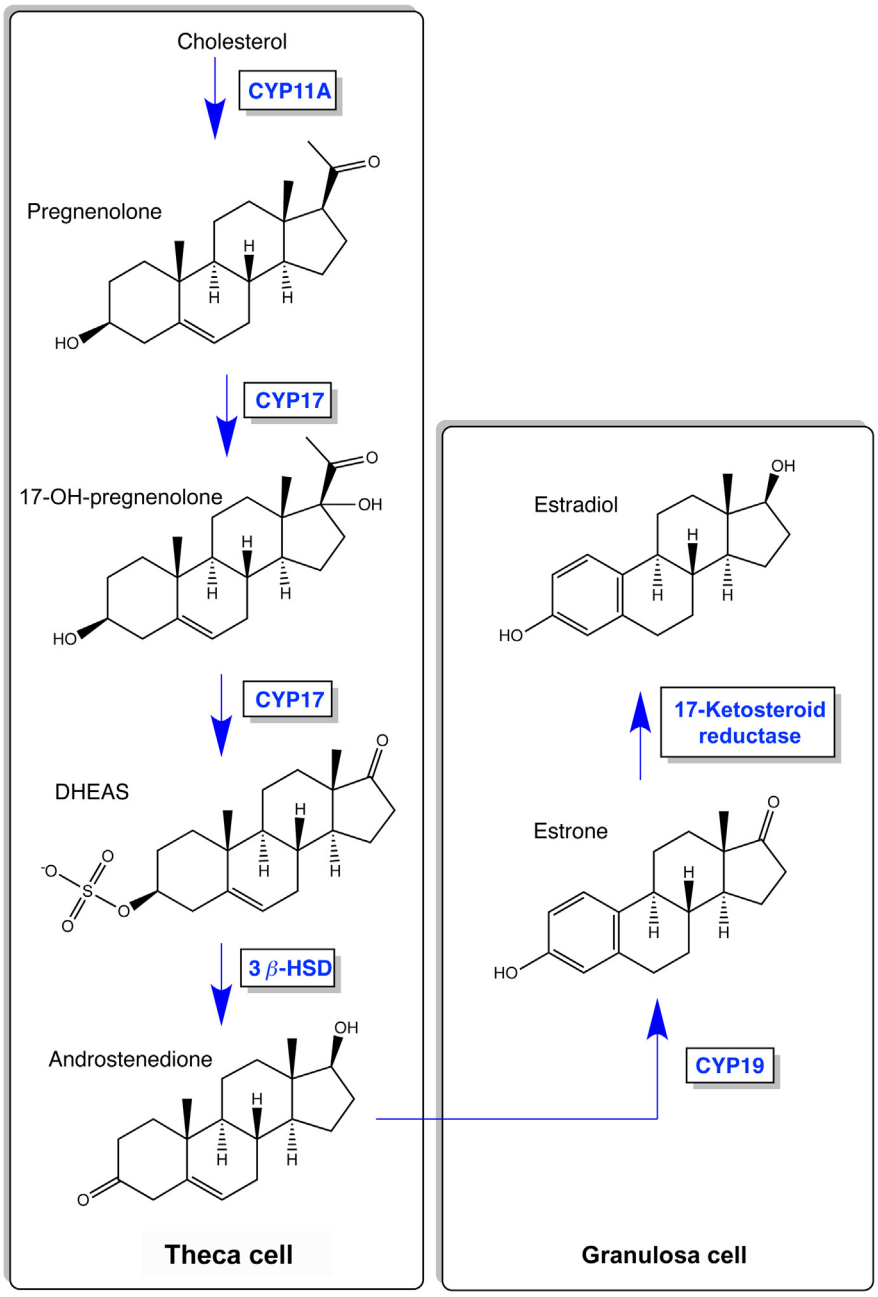


FIGURE 43.4 Estrogen synthesis in granulosa and theca cells that surround oocytes.

Laboratory workup includes measurement of FSH and LH, typically via immunoassay. Concentrations of FSH and LH are typically low with hypothalamic and/or pituitary dysfunction. Follow-up estradiol measurements may help determine the etiology of the hypothalamic/pituitary dysfunction (Table 43.3).

Congenital disorders

A variety of congenital defects can affect reproductive development in women. These range from chromosomal abnormalities, such as Turner syndrome (45, X0), to single-gene mutations in the steroidogenic enzymes, as occur in congenital adrenal hyperplasia (CAH). CAH is caused by deficiency of any of the cytochrome P450 enzymes or 3β-hydroxysteroid dehydrogenase, all of which are required for the synthesis of cortisol and/or aldosterone. The resulting lack of cortisol stimulates pituitary release of ACTH, which drives the steroid pathways causing adrenal hyperplasia. As a consequence, there is a relative excess or deficiency of sex steroids and mineralocorticoids depending on the type of enzyme defect. The most common type of CAH is 21-

TABLE 43.3 Reproductive disorder, etiology, and laboratory results in males and females.

	Abnormality type	Affected sex	Cause	Laboratory results
	Congenital	Females	Turner syndrome	↑FSH, LH ↓Estradiol
		Males	Klinefelter syndrome (XXY) 5α-reductase deficiency Androgen resistance (testicular feminization)	↑FSH, LH ↓Testosterone
	Primary gonadal dysfunction	Females	Menopause Premature ovarian failure Ovarian tumors Chemotherapy or radiation	↑FSH, LH ↓Estradiol
		Males	Trauma Tumors Chemotherapy or radiation Andropause	↑FSH, LH ↓Testosterone
Hypogonadism	Hypothalamic–pituitary dysfunction	Both	Severe weight loss Psychological (eating disorders and stress) GnRH deficiency Cranial irradiation Brain injury Hyperprolactinemia Cushing's syndrome Empty sella syndrome Pituitary infarction	↓FSH, LH ↓Testosterone ↑Prolactin (in cases of prolactinoma)
	Iatrogenic	Both	Prolactin stimulating drugs (reserpine and phenothiazides)	↓FSH, LH ↓Testosterone (male); ↓estradiol (female)
	Physiological	Females	Pregnancy Lactation Lean body weight due to exercise	↑hCG ↑Prolactin ↓FSH, LH ↓Estradiol
Hypergonadism	Congenital	Both	Congenital adrenal hyperplasia	↑Androgens
	Primary gonadal dysfunction	Females	Hormone-secreting tumors • Granulosa cells tumors • Theca cell tumors	↑Total inhibins ↑Estrogen
			Polycystic ovarian syndrome	↑LH/FSH ratio
		Males	Hormone-secreting tumors • Germ cell tumors	↑hCG and/or AFP
	Hypothalamic–pituitary dysfunction	Females	Central ovarian hyperfunction	↑FSH, LH ↑Estradiol
		Males	Hypothalamic hemartoma (ectopic GnRH neurons)	↑FSH, LH ↑Estradiol

GnRH, Gonadotropin-releasing hormone; *FSH*, follicle-stimulating hormone; *LH*, luteinizing hormone.

hydroxylase deficiency, occurring in more than 90% of all cases. This deficiency leads to increased concentrations of 17-hydroxyprogesterone (17-OHP) and adrenal androgens with subsequent virilization. In severe cases (absent 21-hydroxylase) females are born with ambiguous genitalia, whereas in milder forms (nonclassical CAH) women may not undergo sexual maturation. CAH also affects males, who appear phenotypically normal at birth while suffering adrenal insufficiency; newborn screening identifies high 17-OHP concentrations.

A less common form of CAH is 17α-hydroxylase deficiency, an enzyme required for the synthesis of testosterone and estradiol. Deficiency of 17α-hydroxylase causes hypertension and salt wasting due to increased concentrations of deoxycorticosterone (a potent mineralocorticoid), as well as primary amenorrhea (absence of menstruation during adolescence) in females and testicular feminization in males. The laboratory workup for CAH involves measurement of cortisol, aldosterone (to detect mineralocorticoid deficiency), and 17-OHP or other steroid intermediates if 17-OHP is normal. 17-OHP is elevated in classic cases (21-OHase deficiency) and serves as the target analyte in newborn screening tests.

Another congenital cause of infertility in females is Turner syndrome. Turner syndrome results from the absence or dysfunction of one of the X chromosomes. While X-inactivation results in suppression of gene expression from one of the X chromosomes, some genes are not inactivated. Where expression of genes on both X chromosomes is required, XO genotypes cause relative underexpression. This relative genetic underexpression manifests as short stature and sexual infantilism due to a lack of steroid producing follicles at birth. Affected individuals display primary amenorrhea. Identification of a 45, X karyotype is diagnostic for this disorder. Some patients exhibit mosaicism, where only some cells are X0; in these patients, the symptoms may be reduced or absent. Turner patient symptoms are treated with hormonal replacement therapy (e.g., GH and estrogen).

Androgen excess

Affecting 5%−10% of premenopausal women, hirsutism is defined as excessive male-pattern hair growth in woman and is often a sign of androgen excess. Causes of androgen excess include: polycystic ovarian syndrome (PCOS), CAH, ovarian or adrenal androgen-secreting tumors, hyperprolactinemia, and medications. The Endocrine society guidelines suggest measuring plasma androgen concentrations in women with abnormal hirsutism scores (modified Ferriman−Gallwey score) or abnormal hair growth. Androgens in women include testosterone, DHT, androstenedione, DHEA, and DHEA-S. The weakly active adrenal hormones DHEA and DHEA-S can become strong androgens upon conversion to testosterone. Excess androgens in females can result in obesity, acne, and hirsutism (defined as excess hair growth in women in the form of male pattern). The Endocrine Society guidelines recommend against testing androgens in patients with mild hirsutism, as it is unlikely to change treatment. In cases of moderate-to-severe hirsutism, morning testosterone testing is recommended to establish definitive hyperandrogenism. The underlying cause is established through a rule of common symptoms shown in Table 43.2.

PCOS occurs in 8%−13% of women. In women of reproductive age, PCOS is the most common endocrine disorder and most common cause of androgen excess. PCOS presents with a variety of clinical signs including menstrual abnormalities, androgen excess, HPG-axis abnormalities, polycystic ovaries, infertility, obesity, and insulin resistance. Heterogeneity of disease presentation led to disagreement about the definition/diagnostic criteria of PCOS. Recent International Evidence-Based Guidelines for Assessment and Management of PCOS recommend using the Rotterdam criteria to diagnose PCOS in adult women. These criteria require at least two out of the three of the following: ovulatory abnormalities, hyperandrogenism (clinical, biochemical, or both), or polycystic ovarian morphology and the exclusion of related disorders.

The workup for PCOS relies on evidence of androgen excess, menstrual history, and imaging. Imaging of the ovaries is done by ultrasound and reveals the presence of multiple large follicles ($\geq$12) and/or increased ovarian volume. Typically, patients are anovulatory with amenorrhea/oligomenorrhea (established through a careful clinical history and/or low luteal phase progesterone). Hyperandrogenism is established by clinical history and/or clinical exam, and has biochemical evaluation of androgens. Biochemical androgen excess may be identified with total, bioavailable, or free testosterone. Guideline committees have recognized methodological differences, such that there is not a concentration-based cutoff for testosterone for PCOS; guidelines specifically mention preferred use of liquid chromatography−mass spectrometry and extraction chromatography immunoassays for diagnosis. Excess androgens cannot be established in women on oral contraceptives because of the effects on SHBG and gonadotropins. Depending the on the clinical presentation, the laboratory analysis may also include measurement of DHEA-S to identify adrenal androgen excess.

PCOS is currently considered a diagnosis of exclusion, and thyroid dysfunction, hyperprolactinemia, and nonclassical CAH should be ruled out in all adult women with suspected PCOS by measuring TSH, prolactin, and 17-OHP, respectively. In patients with ovulatory dysfunction, pregnancy, hypothalamic amenorrhea, and primary ovarian dysfunction should be ruled out by measuring hCG, FSH, LH, and estradiol. While there is some evidence

supporting anti-Müellerian hormone (AMH) in PCOS, it is not yet recommended for diagnosis because of lack of standardization and robust diagnostic cutoffs. In patients with menstrual abnormalities and/or hirsutism, exclude acromegaly, Cushing's syndrome, and an androgen secreting neoplasm by screening with insulin-like growth factor-1 for acromegaly, one of three tests for Cushing's syndrome (e.g., overnight dexamethasone suppression testing, 24-hour urinary free cortisol, or late night salivary cortisol), and noting a rapid rise and/or significantly elevated testosterone for androgen secreting neoplasm.

Amenorrhea

Clinical history differentiates primary from secondary amenorrhea; primary amenorrhea is the absence of menstruation by adolescence (15−16 years of age), whereas secondary amenorrhea is the absence of menstruation for 6 months in someone who has previously menstruated or its absence for 12 months in those with a history of oligomenorrhea (<9 menstruations/year). Pregnancy is a common cause of secondary amenorrhea. Other causes include premature ovarian failure due to ovarian autoantibodies or prolactinoma. A diagnosis of functional hypothalamic amenorrhea (FHA) is made if organic and anatomical causes are ruled out. FHA is associated with physical or psychological stressors that disrupt normal endocrine function.

Laboratory evaluation of amenorrhea

Measurement of serum LH, FSH, and estradiol is useful in evaluating women with amenorrhea. LH and FSH are increased in primary cases (e.g., genetic causes), but decreased in secondary amenorrhea (e.g., pituitary failure). hCG measurement rules in/out pregnancy, and may be done either in urine or serum/plasma. Prolactin measurement can identify hyperprolactinemia, which causes amenorrhea by disrupting normal pulsatile release of GnRH. The causes of hyperprolactinemia are numerous and include physiological, pharmacological, and pathophysiological causes (see Chapter 39: Disorders of the Anterior and Posterior Pituitary. High TSH concentrations suggest hypothyroidism, which can cause amenorrhea; in primary and secondary hypothyroidism, where thyrotropin-releasing hormone is high, prolactin may be elevated and contribute to amenorrhea. In perimenopause and menopause, waning ovarian function is detected by elevated FSH, decreased estradiol, and slightly increased androgen concentrations.

Assessment of infertility in males and females

A healthy couple attempting to conceive will achieve this at a rate of 20%−25% per month, with 85%−92% occurring within one year. Infertility is defined as the inability to conceive after one year or longer of intercourse without the use of contraception or therapeutic donor artificial insemination. There are numerous causes of infertility in both males and females, necessitating investigation of both partners. Male factors account for 20% of infertility, female factors account for 40%−50%, and 30%−40% of cases are a result of both male and female factors. Endocrine testing and genetic testing (if applicable) are used to identify any of the myriad of causes, which include: anatomical and congenital defects, gonadal failure, malignancies, antisperm antibodies, and psychological, iatrogenic, and endocrine disorders. Guided by the history and physical exam, males typically will have a semen analysis, whereas females will undergo assessment of ovulation and luteal function.

Assessment of infertility in males

Male infertility is divided into four main categories: hypothalamic−pituitary disorders (1%−2%), primary hypogonadism (10%−15%), disorders of sperm transport (10%−20%), and seminiferous tubule dysfunction (60%−80%). Assessment includes a detailed history and physical exam, semen analysis, as well as targeted endocrine and genetic testing.

Semen analysis is considered essential for guiding additional testing to establish a cause of infertility. Though not routinely part of a clinical chemistry menu, semen analysis is subject to the same rigorous regulatory requirements as other testing, and includes quality control and proficiency testing. The Clinical Laboratory Improvement Amendments (and other regulatory bodies in other countries) have defined criteria for semen analysis, and the WHO has defined criteria for volume, pH, viscosity, sperm concentration, count, motility, and morphology. Semen analysis directs subsequent investigations into causes of infertility. Broadly, low sperm counts are divided into pretesticular, testicular, and posttesticular causes. From an endocrine perspective, it is the pretesticular issues that are the focus.

Endocrine evaluation is warranted for men with abnormalities identified by semen analysis (particularly low sperm count) or physical examination findings suggestive of an endocrinopathy. Initially, these findings will prompt measurement of FSH and testosterone at the minimum. Findings of low total testosterone (<300 ng/mL per the ASRM) drive other testing, such as LH and prolactin, to identify primary (gonadal), secondary (pituitary), or tertiary (hypothalamic) causes of infertility. Men with hypothalamic or pituitary failure will have low LH and FSH, which requires a comprehensive investigation and measurement of additional pituitary hormones, such as TSH and prolactin. Men with gonadal failure will display high

FSH and LH with low testosterone. The ability of Leydig cells to produce testosterone can be assessed with an hCG stimulation test. The hCG stimulation test consists of an injection of hCG (5000 IU) with testosterone measurements at baseline and 72-hour postinjection. A healthy response is considered a twofold increase of testosterone from baseline.

Physical examination and imaging may also prompt genetic testing. The congenital bilateral absence of the vas deferens is indicative of a congenital abnormality and a high probability of an abnormal cystic fibrosis transmembrane conductance regulator gene. Karyotypic abnormalities, such as XXY (Klinefelter syndrome), may also be identified as causes of infertility and prompted by physical examination and extremely low sperm count. There are several other rare genetic causes of male infertility, including Y chromosome microdeletions and chromosome translocations. Single-nucleotide polymorphisms in the FSH and LH receptors can also lead to infertility. Collectively, male infertility testing initially involves assessment by a urologist, but may include male infertility specialists, endocrinologists, and clinical geneticists.

Assessment of infertility in females

Female causes of infertility are divided into three categories: anatomic (e.g., congenital abnormalities), ovulatory disorders (e.g., PCOS), and functional problems, such as abnormal cervical mucus. Female assessment of infertility begins with a detailed history and physical examination, which directs the diagnostic evaluation. Investigations may include imaging, biopsy, postcoital testing, ovulation confirmation, endocrine evaluation, and immunological testing.

Approximately 40% of female infertility is the result of ovulatory dysfunction. Ovulatory dysfunction typically manifest as abnormal menstrual cycling. The underlying causes may include PCOS (most common), weight extremes, thyroid dysfunction, and hyperprolactinemia. Ovulatory dysfunction is evaluated with a detailed menstrual history, imaging, biopsy, and endocrine assessment, as well as indirect measurements of ovulation.

Indirect measurements of ovulation are used where the menstrual history does not definitively establish anovulation. Historically, basal body temperature has been used to predict when ovulation occurs, but the American Society for Reproductive Medicine guidelines state that this is not a reliable method to assess ovulatory function, because it does not effectively estimate the timing of ovulation and is tedious for patients. The most accurate assessment of ovulation is a luteal phase progesterone measurement, where a concentration of >3 ng/mL is defined as evidence of ovulation (the ASRM guidelines). Progesterone concentrations begin to rise after ovulation and continue to increase for 5−9 days. In the absence of ovulation, the corpus luteum does not form and cannot secrete progesterone. There are also home testing kits for LH in urine, which can detect the LH peak that occurs 1−2 days prior to ovulation; however, the consistency and reliability of these has been questioned. TSH and prolactin may be measured to establish a treatable cause of anovulation, where thyroid disease and hyperprolactinemia can cause infertility. Imaging may be used to directly assess follicle number and size in complicated cases, whereas endometrial biopsy is now reserved for women with specific endometrial pathology.

In terms of treatment and prognosis, there are a number of laboratory tests that are ordered to assess ovarian reserve. Ovarian reserve is defined by the number and quality of oocytes. Where ovarian reserve is low, so too is the likelihood of fertility. Ovarian reserve is assessed by direct follicle count by ultrasound and/or hormone measurement with or without challenge testing. Clomiphene citrate challenge testing relies on measurement of FSH before and after oral clomiphene citrate (a selective estrogen receptor modulator); elevated postclomiphene FSH levels are consistent with diminished ovarian reserve. More recently, the clomiphene citrate challenge test has been supplanted by measurement of serum AMH. AMH is secreted by granulosa cells of early follicles independent of gonadotropins.

Guidelines state that low serum AMH concentrations (<1 ng/mL, 5.5 pmol/L) predict poor response to ovarian hyperstimulation and low probability of pregnancy success with in vitro fertilization (IVF). Conversely, very high concentrations of AMH (two-to-fivefold higher than the upper reference limit) support a diagnosis of PCOS. Another method for assessing ovarian reserve is "cycle-day 3" serum FSH and estradiol. The prognostic performance of this test is variable, but high FSH on day 3 of the menstrual cycle indicates a low reserve. FSH is elevated when follicles are few and unresponsive, as the hypothalamic−pituitary axis is not suppressed by estradiol. Estradiol is measured to identify if it may be suppressing FSH; high estradiol with low FSH is thus also indicative of low reserve and lower likelihood of successful IVF. Estradiol alone is not effective at assessing ovarian reserve. Overall, female infertility testing is driven by history, where testing is directed based on anatomical and physiological cause factors (Table 43.3).

Suggested reading

W. Winter, et al., Pituitary function and pathophysiology, in: N. Rifai, A. Horvath, C. Whittwer (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, sixth ed., Elsevier, St. Louis, MO, 2018.

R. Nerenz, et al., Reproductive endocrinology and related disorders, in: N. Rifai, A. Horvath, C. Whittwer (Eds.), Tietz Textbook of

Clinical Chemistry and Molecular Diagnostics, sixth ed., Elsevier, St. Louis, MO, 2018.

S.E. Bulun, E.Y. Adashi, The physiology and pathology of the female reproductive axis, in: H.M. Kronenberg, S. Melmed, K.S. Polonsky, P.R. Larsen (Eds.), Textbook of Endocrinology, vol, eleventh ed., Saunders Elsevier, Philadelphia, PA, 2008.

F.A. Conte, M.M. Grumbach, Abnormalities of sexual determination and differentiation, in: F.S. Greenspan, G.J. Strewler (Eds.), Basic and Clinical Endocrinology, vol, fifth ed., Appleton and Lange, Stamford, CT, 1997, pp. 509–546.

D.M. DeKretser, Morphology and physiology of the testis, in: K.L. Becker (Ed.), Principles and Practice of Endocrinology and Metabolism, vol, third ed., Lippincott Williams & Wilkins, Philadelphia, PA, 2001, pp. 1104–1114.

G.F. Erickson, J.R. Schreiber, Morphology and physiology of the ovary, in: K.L. Becker (Ed.), Principles and Practice of Endocrinology and Metabolism, vol, third ed., Lippincott Williams & Wilkins, Philadelphia, PA, 2001, pp. 918–934.

R.A. Lobo, Primary and secondary amenorrhea and precocious puberty: etiology, diagnostic evaluation, management. vol in: V.L. Katz, G. M. Lentz, R.A. Lobo, D.M. Gershenson (Eds.), Comprehensive Gynecology, fifth ed., Mosby Elsevier, Philadelphia, PA, 2007, pp. 933–961.

U.H. Stenman, A. Tiitinen, H. Alfthan, L. Valmu, The classification, functions and clinical use of different isoforms of HCG, Hum. Reprod. 12 (2006) 769–784.

F. Petraglia, J.F. Strauss, S.G. Gabbe, G. Weiss (Eds.), Preterm Birth - Mechanisms, Mediators, Prevention and Interventions, first ed., Informa Heathcare, London, England, 2007.

A. Morales, R.A. Bebb, P. Manjoo, P. Assimakopoulos, J. Axler, C. Collier, , et al.Canadian Men's Health Foundation Multidisciplinary Guidelines Task Force on Testosterone Deficiency, Diagnosis and management of testosterone deficiency syndrome in men: clinical practice guideline, CMAJ 187 (18) (2015) 1369–1377.

R.S. Legro, S.A. Arslanian, D.A. Ehrmann, K.M. Hoeger, M.H. Murad, R. Pasquali, et al., Diagnosis and treatment of polycystic ovary syndrome: an endocrine society clinical practice guideline, J. Clin. Endocrinol. Metab. 98 (12) (2013) 4565–4592.

American Society for Reproductive Medicine, Fertil. Steril. 103 (2015) e18–e25.

American Society for Reproductive Medicine, Fertil. Steril. 103 (2015) e44–50.

World Health Organization, WHO Laboratory Manual for the Examination of Human Semen and Sperm–Cervical Mucus Interaction, third ed., Cambridge University Press, New York, 1992.

K.A. Martin, R.R. Anderson, R.J. Chang, D.A. Ehrmann, R.A. Lobo, M. H. Murad, et al., Evaluation and treatment of hirsutism in premenopausal women: an endocrine society clinical practice guideline, J. Clin. Endocrinol. Metab. 103 (4) (2018) 1233–1257.

C.M. Gordon, K.E. Ackerman, S.L. Berga, J.R. Kaplan, G. Mastorakos, M. Misra, et al., Functional hypothalamic amenorrhea: an Endocrine Society clinical practice guideline, J. Clin. Endocrinol. Metab. 102 (5) (2017) 1413–1439.

S. Bhasin, J.P. Brito, G.R. Cunningham, F.J. Hayes, H.N. Hodis, A.M. Matsumoto, et al., Testosterone therapy in men with hypogonadism: an Endocrine Society clinical practice guideline, J. Clin. Endocrinol. Metab. 103 (5) (2018) 1715–1744.

A. Penzias, K. Bendikson, S. Butts, C. Coutifaris, T. Falcone, G. Fossum, et al., Diagnostic evaluation of sexual dysfunction in the male partner in the setting of infertility: a committee opinion, Fertil. Steril. 110 (5) (2018) 833–837.

H.J. Teede, M.L. Misso, M.F. Costello, A. Dokras, J. Laven, L. Moran, et al., Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome, Fertil. Steril. 110 (3) (2018) 364–379.

A. Sood, D. Cole, F. Abdollah, B. Eilender, Z. Roumayah, M. Deebajah, et al., Endocrine, sexual function, and infertility side effects of immune checkpoint inhibitor therapy for genitourinary cancers, Curr. Urol. Rep. 19 (9) (2018) 68.

Self-assessment questions

1. A 32-year-old woman with a history of irregular periods presents to the fertility clinic after being unable to conceive for 2 years. Physical exam shows an obese woman with excess hair growth on the upper lip and chin, and acne on her chin and back. A pelvic ultrasound shows ovaries with multiple small cysts. What pattern of laboratory tests is consistent with this case?
   a. increased estrogen, increased FSH, and decreased LH
   b. increased testosterone, decreased FSH, and increased LH
   c. increased GnRH and decreased FSH and LH
   d. increased prolactin and decreased TSH
2. A baby girl is born at term with ambiguous genitalia. Evaluation reveals female internal organs on pelvic ultrasonography and a 46, XX karyotype. Newborn screening results show elevated 17-hydroxyprogesterone. Endocrine testing reveals a serum androstenedione of 794 (reference interval 25−95 ng/mL) and a cortisol of 1.2 mcg/dL (reference interval 3−20 μg). Other laboratory values: sodium 128 mmol/L, potassium 6.9 mmol/L, chloride 87 mmol/L, bicarbonate 14 mmol/L, and glucose 40 mg/dL. What is the most likely cause of these findings?
   a. 17α-hydroxylase deficiency
   b. Turner syndrome
   c. 21-hydroxylase deficiency
   d. 5α-reductase deficiency
3. Which hormone is associated with galactorrhea, pituitary adenoma, and amenorrhea?
   a. estradiol
   b. progesterone
   c. FSH
   d. prolactin
4. A deficiency of the enzyme 5α-reductase _____.
   a. results in female pseudohermaphroditism
   b. can be identified by a decreased testosterone:dihydrotestosterone ratio
   c. results in individuals with normal physical but impaired mental development
   d. results in intersex genitalia at birth with eventual masculinization at puberty
5. A possible explanation for an adult male with increased concentrations of the gonadotropins and decreased total testosterone is _____.
   a. primary hypogonadism
   b. secondary hypogonadism
   c. tertiary hypogonadism
   d. complete androgen resistance

Answers

1. b
2. c
3. d
4. d
5. a

Chapter 44

Tumor markers

Lori J. Sokoll and Daniel W. Chan

Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD, United States

Introduction to cancer

In 2020 it is estimated that approximately 1.8 million new cancer cases will be diagnosed and there will be 606,520 deaths from cancer [1]. For new cancer cases, prostate is the leading site in men while breast cancer is the leading site in women. Lung and bronchus cancer, and colon and rectum cancer are the next leading sites for both sexes. Lung and bronchus cancer leads estimated cancer deaths followed by prostate cancer in men and breast cancer in women. Cancer is the second leading cause of death in the United States accounting for one in four deaths; however, the cancer death rate is on a downwards trend, attributable to reductions in smoking and improvements in early detection and treatment.

A simplified definition of cancer is a relatively autonomous growth of tissue. Cancer is a heterogeneous disease with a genetic basis; often driven by changes to growth-promoting protooncogenes, growth-constraining tumor suppressor genes, and DNA repair genes. Causative factors are likely multifactorial including lifestyle factors such as tobacco use, diet, and obesity, exposure to carcinogenic agents such as radiation and chemicals, and nonmodifiable factors including inherited mutations, hormones, and the immune environment. Early detection of cancer affords a greater chance of detecting cancers at an early stage when cancer is organ confined and curative treatment is possible.

Tumors are described in terms of grade and stage. Grade refers to the extent to which a tumor resembles normal differentiated tissue as opposed to poorly differentiated fetal or embryonic tissue. Typical tumor morphology is described as well, moderately, or poorly differentiated. The stage of a tumor describes the anatomic extent of the disease. In the commonly used TNM system (American Joint Commission on Cancer) it is assessed by three components: T (extent of primary tumor), N (regional lymph nodes), and M (metastasis), each of which is followed by a number indicating the extent of malignant disease with the three components combined into an overall stage (0–IV). Clinical staging is based on evidence acquired before treatment, while pathological staging takes into account clinical evidence plus evidence from surgical and pathological examinations. Staging informs prognosis and aids in treatment plans.

Introduction to tumor markers

Definition

A tumor marker is [2]:

- a substance produced by a tumor or by the host in response to a cancer cell;
- found in cells, tissues, or body fluids; and
- measured qualitatively or quantitatively by methods including chemical, immunological, molecular, and mass spectrometry to identify the presence of cancer.

Characteristics

An ideal tumor marker is specific for a given type of cancer, not present in benign conditions, and is released only by or in response to the tumor [3]. The marker should also be sensitive enough to detect small tumors for early diagnosis or during screening and quantitatively reflect response to treatment. Screening requires a marker with high clinical sensitivity and specificity or the marker is used in a population with a high prevalence of cancer. In reality, however, currently available tumor markers are neither specific nor sensitive enough for screening and are used primarily to evaluate progression of disease status after initial therapy and for monitoring subsequent treatment. An approach to improve the utility of tumor markers for screening in a population with low prevalence is to combine markers with other modalities such as combining cancer antigen 125 (CA 125) and ultrasound to detect ovarian cancer. Combining individual markers and other variables through an algorithm into a score or index, such as an in vitro diagnostic multivariate index assay

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00044-2

(IVDMIA), may also improve the utility of tumor markers for a specific intended use. The potential applications of tumor markers are listed in Table 44.1.

Classes

Classes of tumor markers [2] with examples for each class are listed in Table 44.2 and current tumor markers for solid tumors by specimen type presented in Table 44.3. One class of tumor markers is enzymes. Before the discoveries of oncofetal antigens and monoclonal antibodies, enzymes were commonly used as tumor markers with measurements made as enzyme activity. However, with the exception of enzymes such as prostate specific antigen (PSA), which is now measured using antibody-based techniques, most enzymes are not organ-specific. Hormones are another class of tumor markers that have been in use for many years. Commonly used hormones include human chorionic gonadotropin (hCG) for trophoblastic tumors and nonseminomatous testicular cancers, and calcitonin for medullary thyroid cancer. Hormones are also used as tumor markers for gastrointestinal (GI) endocrine tumors such as gastrinomas, glucagonomas, insulinomas, somatostatinomas, vasoactive intestinal polypeptideomas, etc., as well as pituitary tumors such as prolactinomas and adrenal carcinoma [4].

Proteins constitute another class of tumor markers. The first known tumor marker, discovered over 100 years ago, was the Bence–Jones protein, which is still used in multiple myeloma. More recently, proteins and peptides, such as soluble mesothelin-related peptides (SMRP) and progastrin-releasing peptide (proGRP), have been discovered for use as tumor markers in mesothelioma and small cell lung cancer (SCLC), respectively. Oncodevelopmental or oncofetal are the terms used to describe another class of markers that are produced during fetal development and can also be present in increased concentrations due to transformations in malignant cells. Carcinoembryonic antigen (CEA) and α-fetoprotein (AFP) are the most widely used oncofetal antigens.

Carbohydrate markers consist of antigens secreted by tumor cells or present on tumor cell surfaces and include high-molecular-weight mucins and blood group antigens. Discovery of these markers followed the development of monoclonal antibodies and resulted from detection of antigens in tumor cell lines. Examples of these types of tumor markers include CA 125, a marker for ovarian cancer; CA 15-3 and CA 27.29, markers for breast cancer; and CA 19-9, a marker for GI cancers such as colorectal and pancreatic carcinoma. Other tumor marker types include receptors, such as estrogen and progesterone receptors, which are used as prognostic markers in breast cancer patients in addition to predicting response to therapy.

TABLE 44.1 Potential uses of tumor markers.

- Risk stratification
- Screening or early detection
- Differential diagnosis in symptomatic patients
- Clinical staging
- Estimating tumor volume
- Determine prognosis for disease progression
- Monitoring the response to therapy
- Predicting response to therapy
- Detecting cancer recurrence

TABLE 44.2 Tumor marker classification.

Class	Examples
Enzymes, isoenzymes	PSA, LDH, and neuron-specific enolase
Hormones	hCG, calcitonin, ACTH, gastrin, and VIP
Proteins/peptides	β_2-Microglobulin, NMP22, progastrin-releasing peptide, and thyroglobulin
Oncofetal antigens	AFP and CEA
Carbohydrates	CA 125, CA 15-3, and CA 27.29
Blood group antigens	CA 19-9 and CA 72-4
Receptors	Estrogen and progesterone
Gene mutations and overexpression	*BRCA1*, *BRCA2*, and HER-2/*neu*
Other	Circulating tumor cells, and cell-free nucleic acids

ACTH, adrenocorticotropic hormone; AFP, α-fetoprotein; *BRCA*, breast cancer gene; *CA 125*, cancer antigen 125; *CEA*, carcinoembryonic antigen; *hCG*, human chorionic gonadotropin; *HER-2*, human epidermal growth factor receptor-2; *LDH*, lactate dehydrogenase; *NMP*, nuclear matrix protein; *PSA*, Prostate specific antigen; *VIP*, vasoactive intestinal polypeptide.

TABLE 44.3 FDA-reviewed tumor markers for solid tumors in clinical use.

Tumor marker	Associated cancer(s)
Serum and plasma	
CA 15-3, CA 27.29, and HER-2	Breast
CA 125, HE4, OVA1[a], OVERA[a], and ROMA[a]	Ovarian
CA 19-9	Pancreas
Total PSA, Free PSA, cPSA, p2PSA, and prostatic acid phosphatase	Prostate
CEA	Colorectal, breast, and lung
AFP	Nonseminomatous testicular
Thyroglobulin	Thyroid
AFP-L3% (including AFP) and DCP	Hepatocellular
CYFRA 21-1	Lung
β-Human chorionic gonadotropin (for pregnancy, not cleared/approved as a tumor marker)	None
Soluble mesothelin-related peptides	Mesothelioma
Whole blood	
Circulating tumor cells	Breast, colorectal, and prostate
Urine	
NMP22, bladder tumor associated antigens, and UroVysion	Bladder
PCA3	Prostate
Stool	
Hemoglobin (fecal occult blood) and Cologuard (hemoglobin and DNA alterations)	Colorectal
Tissue	
Estrogen and progesterone receptors, MammaPrint, and Prosigna	Breast
P63 antibody	Prostate
Tissue of origin test	Fifteen tumor types

AFP, α-fetoprotein; *CA 19-9*, carbohydrate antigen 19-9; *CEA*, carcinoembryonic antigen; *cPSA*, complexed prostate specific antigen; *CYFRA 21-1*, cytokeratin 19 fragments; *DCP*, des-γ-carboxy prothrombin; *HER-2*, human epidermal growth factor receptor-2; *NMP*, nuclear matrix protein; *p2PSA*, [-2] proPSA; *PCA3*, prostate cancer gene of antigen 3; *PSA*, prostate specific antigen; *ROMA*, risk of ovarian malignancy algorithm.
[a]*In vitro diagnostic multivariate index algorithm.*

Genes or gene products are utilized as tumor markers through determination of mutations or altered gene expression such as human epidermal growth factor receptor-2 (HER-2)/*neu* for the selection of breast cancer patients for Trastuzumab (Herceptin) therapy. Following the classic example of HER-2/*neu*, in the era of personalized or precision medicine, there are now many companion biomarkers (Table 44.4) that predict response or resistance to specific drugs and are used to support treatment decisions.

DNA testing has traditionally been performed on tissue; however, testing of cell-free circulating tumor DNA (ctDNA) in plasma, so-called "liquid biopsy," allows for minimally invasive sampling, can potentially minimize spatial and temporal tumor heterogeneity and DNA sequencing artifacts, and improve turnaround time [5,6]. In 2016 the first plasma DNA test (Roche cobas EGFR Mutation Test v2) was approved by the FDA to aid in selecting targeted therapies for nonsmall cell lung cancer (NSCLC) patients with specific *EGFR* (epidermal growth factor receptor) variants. To address assay sensitivity, testing in corresponding tissue is recommended for negative results. Despite suggested applications for treatment monitoring, residual disease detection, and cancer screening, preanalytical, analytical, and clinical issues currently limit routine use of ctDNA [5,7].

Circulating tumor cells (CTCs) have also been used as tumor markers following methodological development to

TABLE 44.4 FDA cleared or approved tissue or plasma biomarkers for solid tumors used as companion diagnostics (https://www.fda.gov/medicaldevices/productsandmedicalprocedures/invitrodiagnostics/ucm301431.htm).

Gene and or protein biomarkers	Cancer
ALK	NSCLC
BRAF	Melanoma
BRCA	Breast and ovarian
EGFR[a]	NSCLC
EGFR	Colorectal
Estrogen and progesterone receptors	Breast
HER-2; HER-2/*Neu*	Breast and gastric
c-KIT	Gastrointestinal stromal
RAS	Colorectal
PD-L1	NSCLC, gastric, cervical, and urothelial

ALK, anaplastic lymphoma receptor tyrosine kinase; *BRCA*, breast cancer gene; *EGFR*, epidermal growth factor receptor; *HER-2*, human epidermal growth factor receptor-2; *NSCLC*, nonsmall cell lung cancer; *PD-L1*, programmed death-ligand 1.
[a]*Tissue and plasma.*

isolate, identify, and enumerate epithelial cells, with applications for determining disease progression in patients with metastatic breast, prostate, and colorectal cancers. The CellSearch CTC (Menarini Silicon Biosystems) in vitro diagnostic assay uses a combination of positive (epithelial cell adhesion molecule and cytokeratin) and negative (CD45) selection methods to isolate cells from whole blood. In metastatic breast cancer, for example, a CTC count of ≥ 5 cells in 7.5 mL of blood is associated with shorter progression-free and overall survival.

Measurement and evaluation

Tumor markers can be measured qualitatively or quantitatively in body fluids or tissue by numerous methods depending on analyte type. These may include enzyme assays, immunoassays, immunohistochemistry, receptor assays, flow cytometry, or mass spectrometry, as well as molecular techniques. The discovery and analysis of tumor markers as well as their clinical implementation were significantly enhanced by the introduction of the RIA and ELISA techniques as well as the development of monoclonal antibodies in the 1960s and 1970s. Thus the majority of currently used serum and urine tumor markers are measured using immunoassay.

Reference values

Reference ranges for tumor markers can be established in a similar fashion to other analytes from the mean ± 2 SD of large healthy population for a Gaussian (normal) distribution or percentiles (2.5th and 97.5th) for a nonnormal distribution. Typically, however, only the upper cutoff is clinically useful. Another method commonly used to determine cutoff values or decision levels for tumor markers is to use predictive values. Using this approach, patients with the disease are typically compared with patients with benign diseases as opposed to normal subjects. The context in which the biomarker will be used will guide the populations for clinical studies [8]. Cut points can be determined based on the application of the test and requirements to maximize clinical sensitivity, specificity, or both. Using receiver operating characteristic (ROC) curves, plots of sensitivity (true-positive rate) versus 1-specificity (false-positive rate), the range of decision points and corresponding classification can be presented graphically. The area under the curve (AUC) represents overall diagnostic performance, and thus ROC curves are useful for comparing different methods or analytes. A test with 100% sensitivity and 100% specificity would have an AUC of 1.0, while a test with no discriminatory power would have an AUC of 0.5. In the hypothetical example shown in Fig. 44.1, test A, with a greater AUC, has increased clinical performance compared with test B. The cutoff value with the greatest sensitivity and specificity is the point located closest to the upper left-hand corner of the plot.

Analytical considerations

Although the criteria for the analytical evaluation of tumor marker assays are similar to other assays, there are

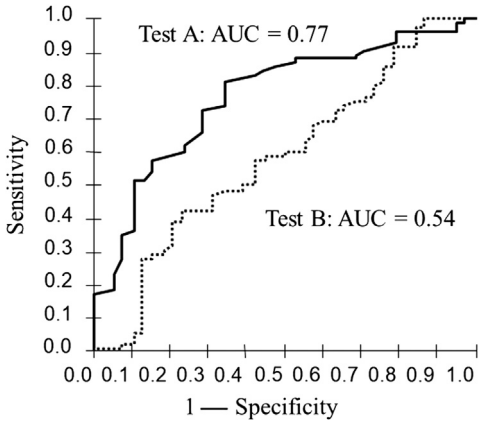


FIGURE 44.1 Receiver operating characteristic curve example.

specific considerations for tumor markers [8,9]. Tumor marker concentrations can span a wide range of values, including undetectable results, and therefore assay linearity is important. When measuring tumor markers, there is potential for a high-dose hook effect (see Chapter 12, Immunoassays) in one-step sandwich assays and for carryover in automated analyzers that do not have disposable sample tips. Heterophilic antibodies and human antimouse antibodies (HAMA) are potential assay interferents although assay manufacturers typically design their assays to minimize HAMA effects. HAMA may also be produced in cancer patients treated with monoclonal antibodies. Test results from patients taking high-dose biotin supplements may be affected in some assays employing biotin−streptavidin designs with false-negative results in sandwich assays and false-positive results in competitive assays.

Interassay precision and lot-to-lot reagent differences are also important since a primary application of tumor markers is monitoring over time. Unlike many assays however, tumor marker assays from different manufacturers may not give the same results, and therefore values are not interchangeable. Typically, reference methods do not exist although international standards and reference materials are available for select analytes (www.nibsc.org). Assay differences may be due to standardization as well as antibody selection and specificity, assay design and kinetics, assay robustness including interferences such as HAMA, as well as variations in reference ranges [10]. When changing assay methods, it may be advisable to offer a crossover period to allow patients being followed to be rebaselined. Including the name of the assay/manufacturer with reported results is also recommended [9].

Clinical considerations

The utility of tumor markers is also evaluated by examining elevations in other disease conditions, both cancerous and noncancerous as well as in the normal population to determine specificity of the test for the type of cancer in question. An example illustrating specificity of risk of ovarian malignancy algorithm (ROMA) for ovarian cancer is shown in Table 44.5.

Tumor markers are most useful for monitoring treatment and progression of cancer [2]. Effective tumor markers should decrease in concentration following initial treatments such as surgery and rise upon disease recurrence following successful treatment. In general, markers used as monitors of therapy should increase with progression, decrease with regression, and remain constant with stable disease. Although there are no standard criteria regarding the relationship between disease status and marker changes, considering analytical and biological marker variability, the following guidelines have been proposed: if no therapy is given, then three consecutive increasing values establish recurrence; if therapy is given, a 25% increase in values indicates progression, a 50% decrease in values indicates partial remission, while complete remission cannot be established using marker values [11].

Specific tumor markers

Despite the large number of potential tumor markers, only a small number have been approved or cleared by the FDA for specific applications (Table 44.3). In contrast, significantly more tumor markers are in use and available commercially worldwide where regulatory requirements are not as rigorous as those in the United States. The FDA review process is described by Fuzery et al. [12]. To follow are the descriptions of primarily protein tumor markers analyzed by immunoassay organized by organ site, primarily focusing on those reviewed by the FDA.

Breast cancer

CA 15-3 and CA 27.29: CA 15-3 and CA 27.29 are tumor markers for breast cancer [2,13]. CA 15-3 and CA 27.29 assays measuring the high-molecular-weight glycoprotein polymorphic epithelial mucin (PEM), also known as episialin, a product of the *MUC1* gene. PEM is shed from the surface of epithelial cells of several tissues, including breast, lung, colon, pancreas, and ovary. PEM is a transmembrane protein (>300 kDa) with a 69 amino acid cellular domain and an extracellular domain consisting of 20 amino acid repeats (1000−2200 amino acids). In each tandem repeat, there are five serine and threonine amino acids that can be glycosylated. In breast cancer, the *MUC1* gene is upregulated with expression 10-fold higher than that in normal cells, and glycosylation is decreased and incomplete exposing epitopes on the core polypeptide backbone, allowing the development of monoclonal antibodies.

TABLE 44.5 Distribution of the risk of ovarian malignancy algorithm in healthy women, and in other malignant and benign conditions for likelihood of malignancy.

		ROMA result interpretation	
	N	Low likelihood (%)	High likelihood (%)
Healthy subjects			
Premenopausal	120	84.2	15.8
Postmenopausal	120	95.0	5.0
Nonovarian malignancies			
Bladder	40	22.5	77.5
Breast	40	35.0	65.0
Endometrial	40	17.5	82.5
GI	39	46.2	53.8
Lung	40	22.5	77.5
Benign conditions			
Gynecological	381	75.3	24.7
Congestive heart failure	40	57.5	42.5
Hypertension	40	70.0	30.0
Pregnancy	38	81.6	18.4

ROMA, risk of ovarian malignancy algorithm.
Source: Adapted from ROMA instructions for use, 2011–09, Fujirebio Diagnostics.

CA 15-3 and CA 27.29 detect overlapping epitopes on the PEM molecule. The CA 15-3 assay consists of two antibodies: 115D8 and DF3. The DF3 detector antibody was generated against a membrane-enriched extract from a human breast cancer metastatic to liver and binds to a sequence of amino acids in the tandem repeat region. Antibody binding is independent of carbohydrate. The 115D8 capture antibody was generated against human milk fat globulin membranes. The antibody does not bind in the tandem repeat region and is carbohydrate-dependent. CA 27.29 assays utilize the B27.29 antibody. This antibody is carbohydrate-independent and was generated against ascites from patients with metastatic breast cancer. The antibody also binds in the tandem repeat region with a peptide sequence that overlaps with the sequence of the DF3 antibody. Similar to other tumor markers, the CA 15.3 (cutoff ~35 U/L) and CA 27.29 (cutoff 37.7 U/L) assays may give different values; however, both assays have similar clinical performance.

The FDA indications for use of CA 15-3 and CA 27.29 are as aids in the management of breast cancer patients for the early detection of recurrence in stage II and stage III patients previously treated and clinically free of disease, and for monitoring response to therapy in patients with stage IV breast cancer. These markers can be elevated in benign liver and breast disease as well as other cancers including pancreatic, lung, ovarian, colorectal, and liver, and therefore are not sensitive enough for early detection. They are prognostic indicators and correlate with the stage of disease and tumor volume and, as suggested by their approved applications, are most useful for detecting residual disease following initial therapy and correlate with disease progression or regression. The American Society for Clinical Oncology (ASCO) clinical guidelines confirm the adjunctive role of CA 15-3 and CA 27–29 as well as CEA in the metastatic setting for monitoring treatment response. Rising markers with a 20%–30% change may indicate treatment failure when measurable disease is not present [13,14].

Gastrointestinal cancers

CEA: CEA was discovered in 1965 as a marker for colorectal cancer [2]. It was subsequently discovered, however, that CEA was not specific for colon cancer or for cancer. The CEA family consists of up to 36 cell-surface glycoproteins with CEA as one of the major proteins. CEA has a mass of 150–300 kDa with 45%–55% carbohydrate.

CEA functions as an adhesion molecule and may be involved in tumor invasion and metastasis. The upper limit of CEA is approximately 3 ng/mL for nonsmokers and 5 ng/mL for smokers with exact values method-dependent.

CEA is elevated in a number of primary cancers including colorectal, breast, pancreatic, liver, lung, gastric, ovarian, and uterine and is elevated in the majority of patients with metastatic liver disease. CEA can be elevated in a number of benign diseases such as cirrhosis, emphysema, rectal polyps, ulcerative colitis, and benign breast disease. Because of these elevations, CEA is not useful for screening. In colon cancer, CEA concentrations are associated with increasing tumor burden and correlate with the stage of disease. Pretreatment concentrations are also indicative of prognosis with increased concentrations associated with the risk of recurrence and development of metastatic disease [15]. The most useful application for CEA in colorectal cancer is in monitoring the clinical course of disease and therapy (Fig. 44.2). CEA measurements are recommended at baseline and at 2–3 months for 3 years following surgery and then every 6 months until 5 years.

CA 19-9: CA 19-9 is a sialylated derivative of the Lewis^{a} blood group antigen. Approximately 5% of the population have the Le^{a-b-} genotype, and therefore do not produce CA 19-9. CA 19-9 is elevated in all GI cancers as well as other adenocarcinomas. It is elevated in 70%–95% of patients with pancreatic cancer and is primarily used to monitor the course of disease, its approved indication. Elevated concentrations can also be seen in benign GI diseases including acute and chronic pancreatitis, cholangitis, and cirrhosis. The initial monoclonal antibody employed in the CA 19-9 assay was developed from a human colon carcinoma cell line, and in some assay designs, the antibody is used for both capture and detection [2]. A cutoff of 37 U/mL is typically used. CA 19-9 immunoassays lack harmonization underscoring the need to monitor patients with the same assay.

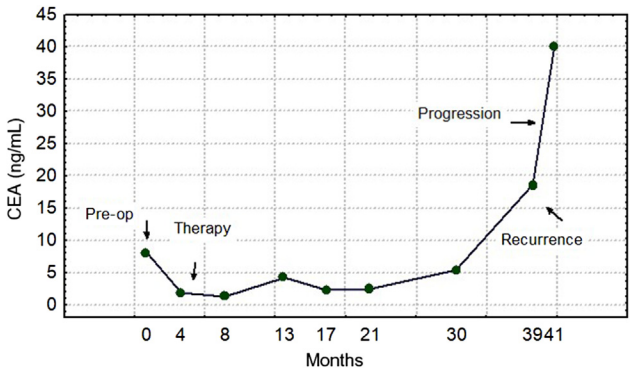


FIGURE 44.2 Carcinoembryonic antigen concentrations for monitoring disease course in a patient with colon cancer.

Genitourinary cancers

Prostate cancer

Prostate specific antigen

PSA is a 28 kDa glycoprotein produced by the prostate epithelium. It is a serine protease and a member of the kallikrein family (hK3) [16]. PSA circulates in the blood in a free form and is bound to protease inhibitors such as α_1-antichymotrypsin (ACT) and α_2-macroglobulin (A2M). PSA is engulfed by the A2M molecule, and therefore PSA-A2M is not measured in total PSA immunoassays. Although PSA is for all intents and purposes organ-specific, it is not cancer-specific and can be elevated in benign prostate diseases such as benign prostatic hyperplasia (BPH) and prostatitis. PSA has replaced prostatic acid phosphatase as a tumor marker for prostate cancer.

Screening for prostate cancer is controversial [17]. Evidence from European randomized trials has shown that PSA screening can reduce metastatic prostate cancer and disease-related death; however, it is acknowledged that screening results in overdiagnosis and overtreatment with potential associated morbidity and mortality. The challenge is to identify men with aggressive disease that will benefit from treatment. It is estimated that approximately half of men diagnosed with prostate cancer have very low-risk disease [18] and are possible candidates for active surveillance. In 2012 the United States Preventive Services Task Force (USPSTF) recommended against PSA screening (grade D), concluding there is more harm from testing than benefit with subsequent decreases in screening, biopsies, and diagnosis of localized disease. In 2018 the recommendations were updated with continued recommendation against screening for men aged 70 or older with acknowledgment that screening may offer a small potential benefit for some men aged 55–69 (grade C), recommending an individualized, informed shared decision-making process between the patient and the physician. Similar to the American Cancer Society and American Urological Association (AUA) recommendations, the USPSTF addresses men at increased risk of prostate cancer including African Americans and men with a family history of prostate cancer.

One of the FDA indications for PSA is to aid in the detection of prostate cancer when used in conjunction with a digital rectal examination (DRE) in men age 50 and older. The associated cutoff of 4 ng/mL has traditionally been used although it is recognized that there is a risk of prostate cancer over all concentrations of PSA [19]. PSA lacks specificity for cancer due to elevations in PSA in men with benign prostate disease, and there is considerable overlap in the PSA range of 4–10 ng/mL, which has been termed the diagnostic gray zone. A number of approaches have been proposed and investigated to

improve the diagnostic accuracy of PSA including age-specific reference ranges, PSA density, PSA velocity, algorithms/nomograms, and the molecular forms of total and free PSA [19]. The concept behind age-specific reference ranges is that by lowering the upper end of the reference range in younger men and extending it in older men, the sensitivity (cancer detection) would be increased in the younger group while specificity (spare unnecessary biopsies) would be increased for the older group. PSA density normalizes PSA concentrations for prostate size using transrectal sonography measurements, while PSA velocity monitors the change in PSA concentrations over time. Nomograms/algorithms look for patterns in variables such as laboratory data and demographic information, and combine information into one index or output measure. An example is the Prostate Cancer Prevention Trial prostate cancer risk calculator (http://myprostatecancerrisk.com/) [20]. Use of the molecular forms of total and free PSA to improve the specificity of PSA, however, has shown the most potential. Biomarkers, including percent free PSA, prostate health index (*phi*), and the 4 kallikrein (4K) score are incorporated into the National Comprehensive Cancer Network (NCCN) prostate cancer early detection guidelines [17] as second line tests for men with specific PSA or DRE results considering first biopsy or repeat biopsy following a negative biopsy.

In the early 1990s it was discovered that the majority of PSA in serum is in the complexed form with only a small portion in the free form and that men with prostate cancer and nonprostate disease differ in proportions of free PSA and PSA bound to ACT. Although the mechanism is unknown, there appears to be a greater percentage of PSA-ACT in patients with prostate cancer than those with benign prostate disease. Measurement of percent free PSA (free PSA/total PSA $\times$ 100) in men with a PSA between 4 and 10 ng/mL is intended to spare biopsies in men with benign disease. Although cutoffs may differ among assays, a recommendation has been made such that if percent free PSA in men with a total PSA between 4 and 10 ng/mL and a negative DRE is greater than 25%, the likelihood of prostate cancer is less than 10%, whereas if percent free PSA is less than 10%, there is a greater than 50% chance of cancer. The test may be most useful in men with an initial negative biopsy.

Subsequently, it was discovered that free PSA consists of the distinct, enzymatically inactive, isoforms benign PSA (BPSA), intact PSA (iPSA), and proPSA [21]. BPSA is a degraded form with two internal peptide bond cleavages found in the transition zone of the prostate where BPH occurs and is associated with benign prostate disease when measured in serum. iPSA is similar to native PSA except that it is enzymatically inactive owing to conformational changes. proPSA is the proenzyme form of PSA with a seven amino acid pro-leader peptide although truncated forms with fewer amino acids exist. proPSA is associated with the peripheral zone of the prostate where the majority of cancers occur. The [-2]proPSA (Beckman Coulter, Inc.) form has been commercialized and is a component of *phi* in combination with total and free PSA. The indication for use is the same as percent free PSA. Ranges of the *phi* index are associated with a probability of prostate cancer on biopsy. A $phi > 35$ is the index defined by the NCCN in men with a PSA of >3.0 ng/mL as informative in the decision for initial or repeat biopsy [17]. The 4Kscore combines PSA, free PSA, iPSA, and hK2 with age, DRE, and prior biopsy status into an algorithm that returns the probability of clinically significant prostate cancer on biopsy [17,20]. Despite inclusion in the NCCN guidelines, the 4Kscore is a lab-developed test (OPKO Health, Inc.) and has not undergone FDA review.

An assay for complexed PSA (cPSA) has been developed that measures PSA-ACT and other PSA complexes irrespective of PSA-A2M. A cPSA concentration of 3.2 ng/mL is equivalent to a PSA cutoff of 4.0 ng/mL, while a PSA concentration of 2.5 ng/mL corresponds to a cPSA concentration of 2.2 ng/mL [22]. cPSA has similar sensitivity for cancer detection compared with total PSA with increased specificity, particularly in the 2−6 ng/mL PSA range. Availability of an automated cPSA assay is limited to one immunoassay manufacturer, Siemens Healthineers.

Serum concentrations of PSA have been found to correlate with tumor volume as well as clinical and pathological stage, with higher levels associated with advanced stages [2]. However, similar to most tumor markers, PSA is most valuable in monitoring the treatment and progression of prostate cancer. PSA can be used to determine the success of initial surgical or radiation treatment. Following radical prostatectomy, PSA concentrations should drop to undetectable levels if the tumor was organ confined and all prostatic tissue has been removed. Detectable PSA levels, determined after a suitable time for concentrations to return to baseline levels, would suggest remaining prostate tissue or the presence of metastases. The PSA cut point for biochemical recurrence defined by the AUA is 0.2 ng/mL. Ultrasensitive PSA and third-generation assays are defined as those with an limit of quantitation (20% CV) of 0.01 ng/mL or lower and many automated assays now achieve this limit. Although cancer recurrence may be detected earlier with lower detection limits, the effect on clinical management, such as for salvage radiation following radical prostatectomy, is unclear and patient-specific. Fourth- and fifth-generation assays with even lower detection limits have been developed although none, including the lab-developed test from Quanterix (AccuPSA) and the FDA-approved NADiA ProsVue assays, are currently available for clinical use.

PSA and its isoforms are measured using sandwich immunoassays. To aid in the standardization of PSA assays, two World Health Organization (WHO) preparations were introduced in 1999. One preparation consists of 100% free PSA, whereas the other consists of 90% PSA-ACT and 10% free PSA. The majority of total PSA assays in use by clinical laboratories are standardized to either the WHO standards or the Beckman Coulter, Inc. Hybritech PSA method. Results from Hybritech standardized assays are approximately 20% higher than WHO standardized assays as a result of different molar absorptivities assigned to PSA.

Prostate cancer gene of antigen 3

Prostate cancer gene of antigen 3 (PCA3) is a molecular urine test for prostate cancer [23]. PCA3 is overexpressed in prostate cancer tissue compared with normal or benign prostate tissue. The PCA3 mRNA is noncoding and has an unknown function. PCA3 is measured in a urine specimen collected after an attentive DRE of three strokes per prostate lobe, which serves to release prostate cells and increase the probability of a sufficient sample. Addition of a stabilization buffer lyses cells and stabilizes the RNA. PCA3 and PSA mRNA are quantitated using transcription-mediated nucleic acid amplification. The PCA3 mRNA copy number is normalized with the PSA mRNA housekeeping gene to generate the PCA3 score (PCA3 mRNA/PSA mRNA $\times$ 1000). The test was approved by the FDA in 2012 (PROGENSA PCA3, Hologic, Inc.) to assist physicians in determining the need for a repeat prostate biopsy in a patient with one or more previous negative biopsies. There is a decreased likelihood of a positive biopsy if the PCA3 score is <25 with a sensitivity and specificity of 77.5% and 57.1% and positive and negative predictive values of 33.6% and 90.0%, respectively. PCA3, with serum PSA and urinary TMPRSS2:ERG (T2:ERG), is part of the lab-developed test from the University of Michigan Mi-Prostate Score for predicting risk of prostate cancer or high-grade prostate cancer on needle biopsy [20,24]. The prostate cancer-specific gene fusion, TMPRSS2:ERG, present in approximately 50% of tumor foci, can be detected in urine using the same methodology as PCA3.

Bladder cancer markers

Urothelial [transitional cell carcinoma (TCC)] bladder cancer presentation includes nonmuscle-invasive, muscle-invasive, and metastatic disease. Seventy-five percent of patients have nonmuscle-invasive disease, which may be treated locally with resection with immunotherapy and/or chemotherapy, while those with muscle-invasive disease initially undergo radical cystectomy with or without perioperative chemotherapy. Bladder cancer has a high rate of recurrence; thus intensive surveillance, generally consisting of cystoscopy and urine cytology, at regular intervals is required. Tumor markers for bladder cancer have been proposed to aid in the management of patients with bladder cancer in order to detect early recurrences of cancer and to complement or replace urine cytology [25].

Several urine tests for TCC of the bladder have been approved for use as aids in the management of bladder cancer patients in conjunction with cystoscopy, including both qualitative and quantitative tests. The bladder tumor antigen *stat* assay is a qualitative test in a point-of-care test format that measures human complement Factor H-related protein (hCFHrp).

The NMP22 assay measures nuclear mitotic apparatus (NuMA), a nuclear matrix protein [2,25]. Nuclear matrix proteins make up the internal structural framework of the nucleus and can function in DNA replication, RNA synthesis, hormone binding, and in regulation of gene expression. NuMA is present in normal and malignant tissues, but can be overexpressed in bladder cancer and, therefore, released into the urine in higher concentrations. Slight elevations above the 10 U/mL cutoff have also been observed in some benign urological conditions as well as in prostate and kidney cancers. In a study of patients resected for TCC, sensitivity of the assay was 70% for detecting recurrence compared with cystoscopy and specificity was 86%. NMP22 assays are available in an ELISA format and an immunochromatographic, point-of-care format (Abbott NMP22 BladderChek) with indications for diagnosis and monitoring.

Hepatocellular cancer

α-Fetoprotein

Another oncofetal antigen, AFP, is a tumor marker for hepatocellular (HCC) and germ cell (nonseminoma) carcinoma [2]. AFP is a 70 kDa single polypeptide chain with 4% carbohydrate. AFP has sequence homology to albumin and is one of the major proteins in the fetal circulation. AFP is produced by the fetal liver and yolk sac with a peak in concentration at 13−14 weeks of gestation. AFP can enter the maternal circulation, and maternal serum concentrations are used in the detection of neural tube defects. AFP concentrations in infants decline to concentrations found in normal adults (<10 ng/mL) by 18 months after birth.

In addition to pregnancy, AFP can be elevated in benign liver conditions such as hepatitis and cirrhosis. A large majority of individuals diagnosed with HCC have cirrhosis with major causes including hepatitis B and C viruses, alcohol, and nonalcoholic fatty liver disease [26]. Ninety-five percent of patients with benign liver disease will have AFP concentrations <200 ng/mL, while

concentrations greater than 1000 ng/mL suggest the presence of cancer in the nonpregnant patient. The American Association for the Study of Liver Diseases recommends surveillance (targeted repeated screening in at risk patients) every 6 months for individuals with cirrhosis using liver ultrasound with or without AFP measurements. Overall survival is improved with the combined modality. The AFP cutoff of 20 ng/mL has a sensitivity of 60% and a specificity of 90%. AFP is also useful in HCC as a prognostic indicator of survival and for monitoring therapy and clinical status.

The glycosylation properties of AFP have been exploited to increase its specificity. Based on the differential binding of AFP to the lectin lens culinaris agglutinin, AFP can be classified into three isoforms: AFP-L1, AFP-L2, and AFP-L3 [27]. AFP-L3 reactivity is associated with liver cancer and AFP-L3%, AFP-L3 as a percentage of AFP, has been developed as a clinical assay with a cutoff of 10% (FUJIFILM Wako Diagnostics). It is indicated for use in patients with chronic liver diseases to assess the risk of developing HCC carcinoma and is prognostic in patients already diagnosed with HCC carcinoma.

AFP, in combination with hCG, is useful for classifying and staging testicular germ cell tumors. Germ cell tumors may be of a single cell type or a mixture of seminomatous and nonseminomatous features. One or both serum markers are elevated in 80%–85% of men with nonseminomatous testicular tumors. In seminomas, hCG is elevated in less than 25% of cases, and in pure seminomas, AFP is not elevated. AFP, hCG, and lactate dehydrogenase have prognostic significance and are used for staging prior to orchiectomy adding on an "S" for serum tumor makers to the traditional T, N, and M staging system [28]. AFP and hCG are most useful for monitoring treatment response and detecting recurrence.

Des-γ-carboxy prothrombin

An additional serum tumor marker for HCC, Des-γ-carboxy prothrombin (DCP) or protein-induced by vitamin K absence or antagonism II (PIVKA-II), is an abnormal form of the coagulation factor prothrombin [27]. In the liver, 10 glutamic acid (glu) residues on prothrombin are posttranslationally modified to γ-carboxy glutamic acid (gla) to form a functional molecule. The γ-glutamyl carboxylase requires vitamin K as a cofactor, and in the cases of dietary deficiency or antagonism, such as by warfarin, DCP is produced. In HCC, gene expression of the enzyme is defective, resulting in DCP.

DCP is most commonly used clinically in countries with a high prevalence of HCC, such as Japan, for early detection, monitoring, and recurrence. In patients with HCC versus those with cirrhosis or chronic hepatitis, sensitivity is 48% and specificity 96%. DCP, and AFP and AFP-L3% are independent markers, and in a case control study of patients with HCC carcinoma or cirrhosis, clinical sensitivity for cancer detection in early stage disease increased to 78% compared with 53% for AFP alone and 61% for DCP alone. DCP correlates with tumor size, although it is less sensitive for detecting small tumors. In the United States, the FUJIFILM Wako microfluidic chip-based immununochemical DPC assay has been FDA cleared as an aid in the risk assessment of patients with chronic liver disease for progression of HCC carcinoma. In a multicenter prospective study in 441 subjects with liver disease, the relative risk of developing HCC carcinoma was 4.8 with a DCP cutoff of 7.5 μg/L.

Thoracic cancer

Lung

The majority of lung cancers are classified histologically as small cell (SCLC), approximately 10%–15% of cases; nonsmall cell (NSCLG), approximately 80%–85% of cases; and, more rarely, carcinoid tumors. NSCLC encompasses adenocarcinoma, squamous cell, and large cell subtypes. A number of nonspecific serum tumor markers have been proposed for lung cancer [29]. Neuron-specific enolase and ProGRP are most associated with SCLC, while CEA and CYFRA 21-1 are most associated with NSCLC. Squamous cell carcinoma antigen is associated with the squamous cell subtype. These markers, in contrast to tissue markers, have less-defined clinical utility and are less likely to be included in clinical guidelines. Availability in the United States is also limited, and only CEA, with a general indication for monitoring in cancer patients, and CYFRA 21-1 are FDA cleared.

CYFRA 21-1

The cytokeratin family consists of approximately 20 proteins that comprise the cytoskeletal intermediate filaments of epithelial cells [2]. Soluble cytokeratin fragments with cancer-associated applications are tissue polypeptide antigen, tissue polypeptide-specific antigen, and cytokeratin 19 fragments (CYFRA 21-1). CYFRA 21-1 is elevated in bladder, lung, and other cancers, and although associated with small cell and NSCLC, it is most sensitive for NSCLC, primarily squamous cell carcinoma. In 2011 an enzyme immunoassay (EIA) for CYFRA 21-1 from Fujirebio Diagnostics was approved for the quantitative determination of soluble cytokeratin 19 fragments in human serum to be used as an aid in monitoring disease progression during the course of disease and treatment in patients with lung cancer. In the 100 patients studied, CYFRA 21-1 serum concentrations reflected change in disease status in 76% of 314 serial specimens. The EIA for CYFRA 21-1 detects cytokeratin 19 fragments with

monoclonal antibodies BM 19.21 and KS 19.1. CYFRA 21-1 concentrations can be elevated in renal disease and other benign conditions and is not affected by smoking status. In 2017 the Elecsys CYFRA 21-1 electrochemiluminescence assay was FDA cleared on Roche immunoassay platforms.

Mesothelioma

Soluble mesothelin-related peptides

SMRP, fragments of the cell-surface glycoprotein mesothelin, can circulate in patients with mesothelial tumors. Mesothelioma is a rare cancer of the mesothelial cells that line the internal organs of the body with the most common form found in the pleura. Mesothelioma is linked to asbestos exposure. The MESOMARK (Fujirebio Diagnostics) assay measures soluble mesothelin-related proteins in an ELISA format and is intended as an aid in monitoring patients diagnosed with epithelioid mesothelioma. A cutoff of 1.5 nM is used, which was derived from the 99th percentile of healthy subjects. It is elevated in 52% of patients with mesothelioma and 10%−15% of patients with ovarian, lung, colon, pancreas, and other cancers. The test has an FDA Humanitarian Device Exemption, which is for medical devices for diseases affecting fewer than 4000 individuals a year; demonstration of test effectiveness is not required.

Ovarian cancer

Cancer antigen 125

Cancer or carbohydrate antigen 125 (CA 125) is a glycoprotein containing 24% carbohydrate with a molecular mass >200 kDa [30,31]. The antigen is so named due to recognition by the OC125 monoclonal antibody. The antibody was developed from a cell line generated from a patient with an ovarian serous papillary cystadenocarcinoma. The antigen is produced by tissues of mullerian duct origin including ovarian epithelium. Secretion of CA 125 is related to the signal transduction pathway of the tyrosine kinase epithelial growth factor receptor.

CA 125 is a tumor marker for epithelial ovarian and endometrial cancers. It is also elevated in pancreatic, lung, breast, and GI cancers as well as in benign conditions such as pregnancy, endometriosis, hepatitis, and others [1,31]. Ovarian cancer is often detected at a late stage with a decreased survival rate compared with early stage disease, and thus sensitive and specific early detection markers are needed. Recent evidence also suggests that high-grade serous carcinoma originates in the fallopian tubes. It is generally accepted that a CA 125 measurement alone is not effective for screening in asymptomatic women, since it is not elevated in early stage disease [19,32]. Approaches to increase the utility of CA 125 for early detection include serial measurements and combining with transvaginal ultrasound. Results from the U.S.US Prostate, Lung, Colorectal, and Ovarian Cancer screening trial showed increased ovarian cancer diagnosis, but did not show a stage shift or reduction in the number of deaths in those randomized to CA 125 and transvaginal ultrasound. The recently reported United Kingdom Collaborative Trial of Ovarian Cancer Screening, which sought to assess ovarian cancer mortality, randomized greater than 200,000 women to no screening, annual transvaginal ultrasound alone, multimodal screening, and triaging with CA 125 testing over time evaluated by risk of ovarian cancer algorithm calculations and ultrasound as a second line test. In the later years of the 14 year study, risk of death was 20% lower in the multimodal screening group and more early stage cancers were detected compared with no screening [31,33]. Early detection in women at high risk for ovarian cancer, such as those with a family history of hereditary disease, is less controversial and screening with CA 125 has been recommended in combination with transvaginal ultrasound [19].

CA 125 may aid in the differential diagnosis of suspicious pelvic masses. A CA 125 > 95 U/L has a high positive predictive value in discriminating malignant and nonmalignant pelvic masses. In the United Kingdom, the risk of malignancy index (RMI) is used for patient management [32]. RMI is calculated from CA 125 concentrations, menopausal status, and ultrasound features. OVA1 is an IVDMIA, the first cleared by the FDA (2009), as aid to assess further the likelihood of malignancy in women with an ovarian adnexal mass with surgery planned who are not yet referred to a gynecologic oncologist. The serum test (Vermillion, Inc.) incorporates CA 125 and four other proteins into a score on a scale of 0−10 with pre- and postmenopausal specific risk cutoffs.

An FDA-approved indication for CA 125 is for monitoring response to therapy for patients with epithelial ovarian cancer. The first approved indication was as an aid in the detection of residual ovarian cancer in previously treated patients considered for diagnostic second-look procedures. However, second-look laparotomy is now considered to be controversial. CA 125 is also useful as a tumor marker for predicting survival pre- and postoperatively and for response to chemotherapy. CA 125 also correlates with disease progression in the majority of cases. The Gynecological Cancer Intergroup has defined response and progression of disease using CA 125. Briefly, a response is described as a 50% decrease in CA 125 from pretreatment that is confirmed and maintained at least 28 days using the same CA 125 assay [34].

The cutoff for CA 125 is 35 U/L. In the initial CA 125 assay, the OC125 monoclonal antibody was used as both capture and detector antibodies owing to multiple binding sites on the antigen. In the CA 125 II assay, the capture

antibody has been replaced with the M11 antibody, which has resulted in improved assay precision.

Human epididymis 4 protein

The human epididymis 4 protein (HE4) is a product of the WAP four-disulfide core domain 2 gene whose function is unknown. The HE4 gene is overexpressed in ovarian cancer, particularly in tumors of serous and endometrioid histology, but is not specific for ovarian tumors. HE4 is also overexpressed in endometrial cancer and as such may be effective as a biomarker [35]. The HE4 protein is measured in serum utilizing 2H5 and 3D8 monoclonal antibodies in a sandwich assay format. The test is cleared by the FDA for monitoring recurrence or progressive disease in patients with epithelial, but not mucinous or germ cell, ovarian cancer. HE4 is available as a manual EIA (Fujirebio Diagnostics) and automated on Abbott, Roche, and Fujirebio (Lumipulse G) automated platforms. HE4 has been incorporated into an algorithm with the CA 125 II assay and menopausal status to aid in assessing likelihood of malignancy in women with an ovarian adnexal mass, an indication similar to OVA1. ROMA was approved by the FDA in 2011. A sensitivity of 93.5% and specificity of 74.9% for ROMA have been reported in 472 women evaluated for an ovarian adnexal mass [35]. The recently FDA cleared second-generation OVA1 test, Overa, incorporates HE4 to improve test specificity.

Thyroid cancer

Thyroglobulin

Thyroglobulin (Tg) is a 660 kDa dimeric glycoprotein synthesized in the follicular colloid of the thyroid gland and is the precursor protein for thyroid hormone synthesis [2]. Tg is regulated by thyroid stimulating hormone (TSH). Serum Tg concentrations are used as a tumor marker to monitor the clinical course of patients with well-differentiated (papillary or follicular) thyroid cancer. Most normal euthyroid individuals have detectable serum Tg concentrations. The majority of patients with elevated Tg concentrations have benign thyroid conditions, since Tg is a nonspecific indicator of thyroid dysfunction. Concentrations can be elevated in Hashimoto's thyroiditis, Graves' disease, thyroid adenoma, and subacute thyroiditis. Therefore Tg is not useful in diagnosing thyroid cancer, and it is used for detecting residual thyroid tissue or recurrent disease.

Tg can be measured by RIA, sandwich immunoassay, or liquid chromatography tandem mass spectrometry (LC-MS)/MS) with immunoaffinity enrichment [36,37]. Second-generation sandwich immunoassays have functional sensitivities of 0.1 ng/mL, allowing monitoring for recurrence postthyroidectomy without the need for recombinant human TSH stimulation. Current RIA and LC-MS/MS assays do not meet the second-generation definition with typical functional sensitivities of 0.5 ng/mL. Tg autoantibodies (TgAb) can cause an interference in Tg immunoassays, most notably in sandwich assays resulting in an underestimation in Tg concentrations. Approximately, a third of thyroid cancer patients have autoantibodies present at some point during their disease and 10% of normal subjects also have TgAb. TgAb should be analyzed with each Tg measurement to aid in the interpretation of results. Similar to other tumor markers, monitoring should be performed with the same Tg assay, irrespective of methodology, facilitating trending, and doubling time calculations. TgAb assay results and cutoffs can also vary among manufacturers and cutoff definition (normal population versus assay sensitivity) can further obscure determination of antibody status. In addition, recognition that trends in TgAb concentrations can reflect response to Tg-secreting thyroid tissue, therefore acting as a tumor marker further emphasizing the need to follow patients with a specific assay as well as the need for assay harmonization [36].

Unlike immunoassays, it is thought LC-MS/MS Tg methods are devoid from antibody interferences although these methods are less well-characterized for their ability to assess accurately disease status in thyroid cancer patients. Mass spectrometry methods may be incorporated into reflex strategies for Tg testing, whereby initial TgAb dictates triage to immunoassay in TgAb-negative specimens and mass spectrometry or RIA in TgAb positive specimens.

Another valuable application for Tg measurement is as an aid to cytology in detecting lymph node metastases in patients treated for differentiated thyroid cancer with total thyroidectomy [38]. Cytomorphologic examination can be nondiagnostic, most often due to insufficient cellularity. Following ultrasound-guided fine needle aspiration of the suspicious lymph node, the needle is rinsed with saline or buffer and the sample tested for Tg. Detectable Tg is suggestive of the presence of metastatic disease; however, there is no uniform diagnostic cutoff for a negative result, as sample collection methods and assays are not standardized.

Guidelines for the use of tumor markers

As cited above, a number of clinical oncology and laboratory organizations develop and publish guidelines, addressing the clinical use of tumor markers. Clinical practice guidelines from The ASCO (https://www.asco.org/practice-guidelines/quality-guidelines/) and The AUA (https://www.auanet.org/guidelines) focus on specific cancers, while the NCCN (www.nccn.org) develops clinical practice guidelines for diagnosis and treatment of a wide range of cancers.

Laboratory medicine contributions include the Laboratory Medicine Practice Guidelines from The National Academy of Clinical Biochemistry (now AACC Academy: www.aacc.org/community/aacc-academy/publications) and publications from the European Group on Tumor Markers (www.egtm.eu), both addressing clinical as well as analytical and quality considerations.

References

[1] American Cancer Society, Cancer Facts & Figures 2020, American Cancer Society, Atlanta, GA, 2020, pp. 1–76.

[2] L.J. Sokoll, A.J. Rai, D.W. Chan, Tumor markers, in: C.A. Burtis, E.R. Ashwood, D.E. Bruns (Eds.), Tietz Textbook of Clinical Chemistry, fifth ed., W.B. Saunders Company, Philadelphia, PA, 2012, pp. 617–667.

[3] M.J. Duffy, Tumor markers in clinical practice: a review focusing on common solid cancers, Med. Princ. Pract. 22 (2013) 4–11.

[4] L.J. Sokoll, D.W. Chan, Malignancy-associated endocrine disorders, in: W.E. Winter, L.J. Sokoll, I. Jialal (Eds.), Handbook of Diagnostic Endocrinology and Metabolism, AACC Press, Washington, DC, 2008, pp. 401–431.

[5] M.J. Duffy, Circulating tumour DNA as a cancer biomarker, Ann. Clin. Biochem. 56 (2019) 42–48.

[6] H. Husain, V.E. Velculescu, Cancer DNA in the circulation. The liquid biopsy, JAMA 318 (2017) 1272–1273.

[7] J.D. Merker, G.R. Oxnard, C. Compton, M. Diehn, P. Hurley, A.J. Lazar, et al., Circulating tumor DNA in patients with cancer: American Society of Clinical Oncology and College of American Pathologists joint review, J. Clin. Oncol. 36 (2018) 1631–1641.

[8] M.J. Duffy, C.M. Sturgeon, G. Soletormos, V. Barak, R. Molina, D.F. Hayes, et al., Validation of new cancer biomarkers: a position statement from the European group on tumor markers, Clin. Chem. 61 (2015) 809–820.

[9] C.M. Sturgeon, B.R. Hoffman, D.W. Chan, et al., National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines for use of tumor markers in clinical practice: quality requirements, Clin. Chem. 54 (2008) e1–e10.

[10] C.M. Sturgeon, J. Seth, Why do immunoassays for tumour markers giver differing results? - a view from the UK national external quality assessment schemes, Eur. J. Clin. Chem. Clin. Biochem. 34 (1996) 755–759.

[11] J.M.G. Bronfrer, Working group on tumor marker criteria (WGTMC), Tumor Biol. 11 (1990) 287–288.

[12] A.K. Fuzery, J. Levin, M.M. Chan, D.W. Chan, Translation of proteomic biomarkers into FDA approved cancer diagnostics: issues and challenges, Clin. Proteom. 10 (2013) 1–14.

[13] M.J. Duffy, E.W. McDermott, J. Crown, Blood-based biomarkers in breast cancer: from proteins to circulating tumor cells to circulating tumor DNA, Tumor Biol. 40 (2018) 1–11.

[14] C. Van Poznak, M.R. Somerfield, R.C. Bast, M. Cristofanilli, M. P. Goetz, A.M. Gonzalez-Angulo, et al., Use of biomarkers to guide decisions on systemic therapy for women with metastatic breast cancer: American Society of Clinical Oncology Clinical Practice Guideline, J. Clin. Oncol. 33 (2015) 2695–2704.

[15] M.J. Duffy, R. Lamerz, C. Haglund, A. Nicolini, M. Kalousova, L. Holubec, et al., Tumor markers in colorectal cancer, gastric cancer and gastrointestinal stromal cancers: European group on tumor markers 2014 guidelines update, Int. J. Cancer 134 (2014) 2513–2522.

[16] L.J. Sokoll, D.W. Chan, PSA: its discovery and biochemical characteristics, Urol. Clin. North. Am. 24 (1997) 253–259.

[17] National Comprehensive Cancer Network. Prostate cancer early detection (Version 2.2018).

[18] National Comprehensive Cancer Network. Prostate cancer (Version 4.2018).

[19] C.M. Sturgeon, M.J. Duffy, U.H. Stenman, et al., National Academy of Clinical Biochemistry laboratory medicine practice guidelines for use of tumor markers in testicular, prostate, colorectal, breast, and ovarian cancers, Clin. Chem. 54 (2008) e11–e79.

[20] S. Loeb, H. Dani, Whom to biopsy. Prediagnostic risk stratification with biomarkers, nomograms, and risk calculators, Urol. Clin. N. Am. 44 (2017) 517–524.

[21] S.D. Mikolajczyk, L.S. Marks, A.W. Partin, H.G. Rittenhouse, Free prostate-specific antigen in serum is becoming more complex, Urology 59 (2002) 797–802.

[22] A.W. Partin, M.K. Brawer, G. Bartsch, et al., Complexed prostate specific antigen improves specificity for prostate cancer detection: results of a prospective multicenter clinical trial, J. Urol. 170 (2003) 1787–1791.

[23] J.J. Tosoian, A.E. Ross, L.J. Sokoll, A.W. Partin, C. Pavlovich, Urinary biomarkers for prostate cancer, Urol. Clin. N. Am. 43 (2016) 17–38.

[24] S.A. Tomlins, J.R. Day, R.J. Lonigro, D.H. Hovelson, J. Siddiqui, L.P. Kunju, et al., Urine TMPRSS2: ERG plus PCA3 for individualized prostate cancer risk assessment, Eur. Urol. 70 (2016) 45–53.

[25] A. Mbeutcha, I. Lucca, R. Mathieu, Y. Lotan, S.F. Shariat, Current status of urinary biomarkers for detection and surveillance of bladder cancer, Urol. Clin. N. Am. 43 (2015) 47–62.

[26] J.A. Marrero, L.M. Kulik, C.B. Sirlin, A.X. Zhu, R.S. Finn, M.M. Abecassis, et al., Diagnosis, staging, and management of hepatocellular carcinoma: 2018 practice guidance by the American Association for the Study of Liver Diseases, Hepatology 68 (2018) 723–750.

[27] D. Li, S. Satomura, Biomarkers for hepatocellular carcinoma (HCC): an update, Adv. Exp. Med. Biol. 867 (2015) 179–193.

[28] Testis, in: S. Edge, D.R. Byrd, C.C. Compton, et al. (Eds.), AJCC Cancer Staging Handbook, seventh ed., Springer, New York, 2010.

[29] M.J. Duffy, K. O'Byrne, Tissue and blood biomarkers in lung cancer: a review, Adv. Clin. Chem. 86 (2018) 1–21.

[30] I.M. Shih, L.J. Sokoll, D.W. Chan, Ovarian cancer, in: E.P. Diamandis, H.A. Fritsche, H. Lilja, D.W. Chan, M.K. Schwartz (Eds.), Tumor Markers. Physiology, Pathobiology, Technology, and Clinical Applications, AACC Press, Washington, DC, 2002, pp. 239–252.

[31] K.M. Elias, J. Guo, R.C. Bast, Early detection of ovarian cancer, Hematol. Oncol. Clin. N. Am. 32 (2018) 903–914.

[32] G. Soletormos, M.J. Duffy, S. Othman Abu Hassan, R.H.M. Verheijen, B. Tholander, R.J. Bast, et al., Clinical use of cancer biomarkers in epithelial ovarian cancer. Updated guidelines from the European group on tumor markers, Int. J. Gynecol. Cancer 26 (2016) 43–51.

[33] I.J. Jacobs, U. Menon, A. Ryan, A. Gentry-Maharaj, M. Burnell, J. K. Kalsi, et al., Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomized controlled trial, Lancet 387 (2016) 945–956.

[34] G.J. Rustin, I. Vergote, E. Eisenhauer, E. Pujade-Lauraine, M. Quinn, T. Thigpen, et al., Definitions for response and progression in ovarian cancer clinical trials incorporating RECIST 1.1 and CA 125 agreed by the gynecological cancer intergroup (GCIG), Int. J. Gynecol. Cancer 21 (2011) 419–423.

[35] A.R. Simmons, K. Baggerly, R.C. Bast, The emerging role of HE4 in the evaluation of advanced epithelial ovarian and endometrial carcinomas, Oncology 27 (2013) 548–556.

[36] C. Spencer, J. LoPresti, S. Fatemi, How sensitive (second generation) thyroglobulin measurement is changing paradigms for monitoring patients with differentiated thyroid cancer, in the absence or presence of thyroglobulin antibodies, Curr. Opin. Endocrinol. Diabetes Obes. 21 (2014) 394–404.

[37] A.N. Hoofnagle, M.Y. Roth, Clinical review: improving the measurement of serum thyroglobulin with mass spectrometry, J. Clin. Endocrinol. Metab. 98 (2013) 1343–1352.

[38] Q.K. Li, S.L. Nugent, J. Straseski, D. Cooper, S. Riedel, F.B. Askin, et al., Thyroglobulin measurements in fine-needle aspiration cytology of lymph nodes for the detection of metastatic papillary thyroid cancer, Cancer Cytopath 121 (2013) 440–448.

Self-assessment questions

1. Tumor markers _____.
   - **a.** must be produced by the tumor
   - **b.** must be measured in body fluids
   - **c.** must be measured by immunoassay
   - **d.** must be quantitative
   - **e.** none of the above
2. Current tumor markers are most useful for _____.
   - **a.** screening
   - **b.** diagnosis
   - **c.** monitoring disease course
   - **d.** risk assessment
   - **e.** prognosis
3. ROC curves are constructed for two tumor markers for pancreas cancer using patients with cancer compared with normal subjects. The AUCs are 0.75 for test A and 0.54 for test B. What can be concluded about these two tests?
   - **a.** Test A has a better overall clinical performance compared to test B.
   - **b.** Test A is better than CA 19-9 as a tumor marker for pancreas cancer.
   - **c.** These tests are not elevated in patients with benign GI diseases
   - **d.** There is an assay cutoff for test A that provides 100% clinical sensitivity and specificity
   - **e.** All of the above
4. The CA 15-3 breast cancer marker may be elevated in patients with _____.
   - **a.** breast cancer
   - **b.** ovarian cancer
   - **c.** liver disease
   - **d.** lung cancer
   - **e.** all of the above
5. CA 19-9 is analyzed in a patient diagnosed with pancreas cancer prior to surgery. The result is undetectable. Which of the following are true?
   - **a.** The patient does not have pancreas cancer.
   - **b.** The patient has HAMA.
   - **c.** CEA should also be measured.
   - **d.** The marker is not informative for monitoring treatment.
   - **e.** All of the above.
6. A 65-year-old man is referred to a urologist with a PSA of 5.6 ng/mL. His DRE is negative. A 12-core biopsy is performed and no cancer is found. Six months later, his total PSA is 5.4 ng/mL. A percent free PSA is performed. The result is 29%. What can be concluded from this information?
   - **a.** The likelihood of cancer is high.
   - **b.** The likelihood of cancer is low.
   - **c.** This patient has BPH.
   - **d.** This patient has prostatitis.
   - **e.** b and c.
7. Total PSA assays measure _____.
   - **a.** unbound PSA
   - **b.** PSA-ACT
   - **c.** PSA-A2M
   - **d.** a and b
   - **e.** a–c
8. Which of the following is not an approach to improve the clinical utility of PSA for prostate cancer detection?
   - **a.** PSA velocity
   - **b.** Ultrasensitive PSA
   - **c.** Molecular forms of PSA
   - **d.** Molecular forms of free PSA
   - **e.** PSA density
9. Which of the following is true of CA 15-3 and CA 27.29 assays:
   - **a.** Assays cannot be used interchangeably for monitoring disease.
   - **b.** Assays give the some values for a given specimen.
   - **c.** Assays use the same cutoff values.
   - **d.** Assay design and kinetics are similar.
   - **e.** Assays are approved for use in the early detection of breast cancer.
10. Which of the following is/are true for the ROMA and OVA1 algorithms:
    - **a.** They are used to diagnose ovarian cancer.
    - **b.** CA 125 is one of the analytes in the panel of markers.
    - **c.** They are indicated for use only in postmenopausal women.
    - **d.** They have similar indications for use.
    - **e.** b and d.

Answers

1. e
2. c
3. a
4. e
5. d
6. b
7. d
8. b
9. a
10. e

Chapter 45

Calcium biology and disorders

William E. Winter[1,2], J. Peter R. Pelletier[1,3], and Neil S. Harris[1]

[1]Department of Pathology, University of Florida, Gainesville, FL, United States, [2]Department of Pediatrics, University of Florida, Gainesville, FL, United States, [3]Department of Anesthesia, University of Florida, Gainesville, FL, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Explain the regulation of calcium and phosphate.
- List the differential diagnosis of hypocalcemia and hypercalcemia.
- Diagnose conditions involving excesses or deficiencies of calcium.
- Describe the role of calcitonin testing in the diagnosis and monitoring of medullary thyroid carcinoma.
- Choose markers of bone turnover in the evaluation of osteoporosis.

Introduction

In the body, calcium has a major structural role in the formation of bone and is important in blood coagulation, neurological and neuromuscular function, and intracellular signaling. The majority of calcium in the body is localized to bone (e.g., ~99%). Phosphate is involved in many critically important biochemical processes, including energy metabolism (e.g., adenosine triphosphate, ATP), nucleic acid metabolism, cell signaling, bone formation, and maintenance of acid/base balance (especially related to urinary phosphate buffering). Both calcium and phosphate are important in skeletal formation and turnover. Bone provides us (as humans) with form, mobility, and protection. It is also the site of hematopoiesis. In many ways, bone is an extended endocrine target organ, as many hormones regulate bone dynamics to varying degrees, including parathyroid hormone (PTH), 1,25-dihydroxyvitamin D (1,25-OH_2D), calcitonin, thyroid hormones, sex steroids, glucocorticoids, growth hormone, and insulin-like growth factor I. Bone is also the source of many regulatory factors, including fibroblast growth factor-23 (FGF23). The biological action of FGF23 is to produce phosphaturia. FGF23 is produced by osteocytes.

Calcium

Calcium function and distribution

As noted above, the majority of the body's calcium (e.g., ~99%) is localized to bone. Of the remaining 1% of calcium that is not associated with bone, there are two pools of calcium, which are stratified into a slowly exchangeable pool and a rapidly exchangeable pool. The slowly exchangeable pool encompasses calcium in dystrophic sites such as atheromas and damaged cartilage, and in subcellular organelles such as the mitochondria and the endoplasmic reticulum. The rapidly exchangeable pool includes calcium in systemic circulation, interstitial fluid, within cellular cytoplasm, and calcium that has recently been deposited on bone surfaces. Plasma total calcium (not associated with bone) is categorized in the body as: (1) free calcium; (2) protein-bound calcium; and (3) complexed (anion-bound) calcium. Within the circulation, ~40%–50% of calcium is ionized (free). Bound calcium is carried by plasma proteins (predominantly albumin; ~40% of the total calcium) or complexed to anions such as bicarbonate (HCO_3^-), citrate, lactate, and phosphate (~10% of the total calcium).

Ionized calcium is measured using an ion-selective electrode (ISE). The free (unbound) calcium fraction, which is responsible for the electrophysiological actions of calcium, is monitored by the parathyroid gland's cell-surface, transmembrane calcium-sensing receptor (CaSR). The proportion of calcium that is ionized versus bound to plasma proteins is influenced by blood pH. In acidotic states, ionized calcium increases as calcium is "released" by plasma proteins, whereas alkalosis decreases ionized calcium as calcium is bound to plasma proteins. Hyperventilation, for example, through the induction of alkalosis, can produce symptoms of hypocalcemia (e.g., perioral paresthesia and tingling). Ultrafilterable calcium

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00045-4

is the sum of the ionized calcium and anion-complexed calcium that can pass through the glomerular basement membrane of the nephron into the Bowman space. This calcium (and phosphate as well) will be reabsorbed by the renal tubules limiting calciuria (which if excessive can cause renal stones or nephrocalcinosis). Total body phosphate is controlled predominantly by the excretion of phosphate.

In modern laboratories, total calcium is measured using dye-binding assays (e.g., arsenazo III or o-cresolphthalein complexone). Prevention of interference by magnesium is achieved by the addition of 8-hydroxyquinoline to the reaction. Newer generation calcium assays utilize 5-nitro-5′methyl-BAPTA as a substrate for the calcium measurement. On certain analyzers, calcium is measured by ISE after acidification to release protein-bound and complexed calcium into the ionized state. This is important, as ISEs do not show interference from gadolinium, whereas dye-binding methods do display transient negative interferences with some gadolinium formulations [1]. Gadolinium is a contrast agent used in radiology.

Because total calcium levels can be significantly affected by albumin concentrations, if the total calcium is abnormal, either serum albumin or ionized calcium should be measured. Total calcium can be corrected for deficiencies or excesses of albumin as follows: for every decline in albumin of 1 g/dL below 4 g/dL, the total calcium is expected to decline by 0.8 mg/dL assuming there are no disorders of calcium homeostasis. Likewise, for every increase in albumin above 4 gm/dL, the total calcium is expected to rise by 0.8 mg/dL.

$$\text{Corrected total } Ca^{++} = \text{measured total } Ca^{++} + 0.8 * \left[4.0 - \text{measured albumin}(g/dL)\right]$$

Recognizing that these rules are approximations, measurement of ionized calcium is preferred.

Calcium sensing by the parathyroid glands

Ionized calcium is monitored by the CaSR located on the cell surfaces of the parathyroid glands [2] (Fig. 45.1). Four parathyroid glands are normally located behind the thyroid gland. However, the number of parathyroid glands can vary between three and five and ectopic parathyroid tissue can be located elsewhere in the neck, mediastinum, or among the aortic arch. Reduced stimulation of the CaSR because of a low interstitial ionized calcium concentration stimulates the receptor to engage processes that release PTH from the parathyroid cells. On the other hand, increased stimulation of the CaSR because of a high interstitial ionized calcium concentration reduces parathyroid gland PTH release. Although it was initially postulated that the CaSR would function as an ion channel, molecular analysis of the CaSR revealed that this 1078−amino-acid receptor was a member of subfamily C of the G-protein-coupled receptors. Other members of this subfamily include metabotropic glutamate receptors, gamma-aminobutyric acid type B receptors, and pheromone receptors. The CaSR has a large extracellular domain for ionized calcium binding, seven transmembrane domains, and an intracellular domain that couples the receptor to G proteins. The gene for the CaSR is located on chromosome 3q13.3-q21.

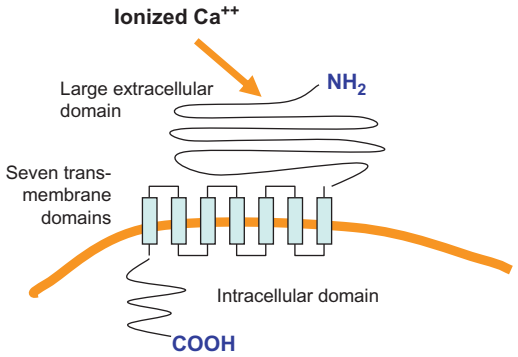


FIGURE 45.1 The calcium sensing receptor has a large N-terminal extracellular domain that binds ionized calcium, a 7-member transmembrane domain, and an intracellular domain.

Parathyroid hormone synthesis, release, and metabolism

A decline in ionized calcium concentrations elicits PTH secretion from the parathyroid glands. The PTH gene is located on chromosome 11p. In contrast, elevations in ionized calcium suppress PTH release. PreproPTH is 115 amino acids in length (Fig. 45.2). Upon entry of preproPTH into the rough endoplasmic reticulum, the 25 amino acid presequence (or leader/signal sequence) is cleaved, yielding proPTH of 90 amino acids. ProPTH is further cleaved to PTH with the removal of another N-terminal, 6 amino acids. This full-length 84−amino-acid form of PTH is termed "1−84 PTH." Either within the parathyroid or shortly after secretion, 6 N-terminal amino acids are removed from ∼50% of the 1−84 PTH molecules, yielding 7−84 PTH within the systemic circulation. In normal individuals, the ratio of 1−84 PTH to 7−84 PTH is ∼1:1. Whereas 1−84 PTH acts to raise calcium concentrations, 7−84 PTH may actually lower calcium concentrations. 1−34 PTH may also be derived from 1−84 PTH [3]. 1−34 PTH also appears to raise calcium concentrations.

In terms of PTH immunoassays, the intact PTH double-antibody (immunometric) assays detect both 1−84 PTH and 7−84 PTH through a standard antibody−antigen−antibody "sandwich" approach. Thus the

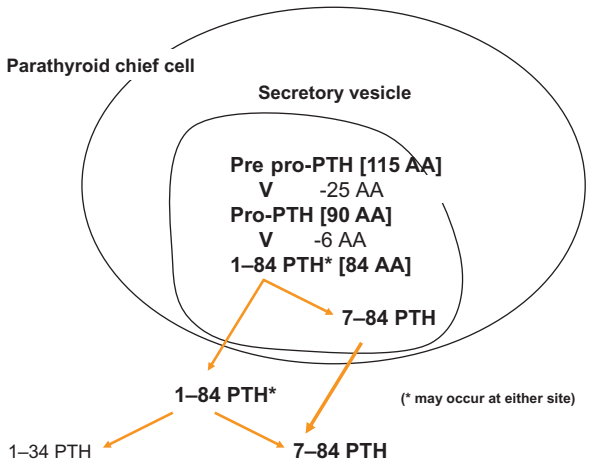


FIGURE 45.2 PTH is initially synthesized as preproPTH (115 amino acids) that is processed to proPTH (90 amino acids) with removal of the pre- or leader sequence (25 amino acids). Removal of a further six amino acids yields 1−84 PTH. Either within parathyroid secretory vesicles or after secretion, the N-terminal 6 amino acids can be cleaved from 1−84 PTH producing 7−84 PTH. A 1−34 PTH fragment can also be produced from 1−84 PTH.

term "intact PTH" assay is really a misnomer, because the intact assay does not exclusively detect 1−84 PTH. In such intact PTH assays, a monoclonal antibody to the N-terminus first recognizes a motif starting at approximately amino acid 7, and thus amino acids 1−6 are not necessary for this assay to detect PTH. A polyclonal antiserum is directed at C-terminus motifs distal to amino acid 34 in PTH.

In response to the recognition that the intact assay does not measure exclusively 1−84 PTH, in the past, several manufacturers developed immunoassays that detected 1−84 PTH but not 7−84 PTH [4,5]. These assays have been variously named "true-intact," "whole," "bio-intact," or "cyclase-activating" PTH assays. These 1−84 PTH assays use an antibody to the N-terminus that binds to the first amino acid in PTH, allowing detection of 1−84 PTH. However, at least some of the whole PTH assays also detect a 1−34 PTH molecule, and thus these assays may not be absolutely specific for 1−84 PTH.

Some investigators have suggested that "whole" PTH assays were superior to intact PTH assays in the assessment of secondary hyperparathyroidism of renal failure [6−12]. Because the intact PTH assays include measurements of 1−84 PTH and 7−84 PTH, a hypothesis was advanced that the whole PTH assays would be superior to the intact PTH assays because of the selectivity of the whole PTH assays for 1−84 PTH. Furthermore, whole PTH assays would not overestimate PTH as would the intact PTH assays. Theoretically, overestimation of the PTH concentration could lead to clinical overtreatment of secondary hyperparathyroidism. Close review of the literature, however, does not support the superiority of the whole PTH assays over the intact PTH assays in this regard. Compared with the intact PTH assay, the whole PTH assay is not a better predictor of bone loss. Furthermore, the ratio of 1−84 PTH to non-1−84 PTH does not correlate with the rate of bone formation.

Presently (November 2018), none of the major reference labs (Mayo Medical Laboratories, LapCorp, Esoterix, Quest, and ARUP) appears to offer true-intact (biointact) PTH assays. Several papers from 2001−05 showed that the correlation of 7−84PTH with 1−84PTH in: (1) primary hyperparathyroidism; (2) uremia; and (3) pediatric dialysis displayed r values of $= > 0.92$. Bone formation rates correlate equally well with 7−84PTH measurements and 1−84PTH measurements regardless of the presence or absence of uremia. These studies lead to the conclusion that 1−84PTH measurements are not clinically superior to 7−84PTH measurements.

Parathyroid hormone actions

Renal tubular cells and osteoblasts respond to PTH via expression of a receptor for PTH. Besides binding PTH, the PTH receptor also binds PTH-related peptide (PTHrP) and is named "PTHR1" or the "PTH/PTHrP" receptor [13]. PTHrP is released by some cancers and can cause hypercalcemia (i.e., humoral hypercalcemia of malignancy). PTHrP is not detected in PTH immunoassays. Intact PTH is 84 amino acids in length versus PTHrP, which is 131 amino acids. PTHrP only shares 8 of its first 13 amino acids with PTH indicating that these proteins are quite different structurally. The N-terminal 14 amino acids in PTH and PTHrP are necessary for the signaling functions of these molecules. There is a separate immunoassay for PTHrP. PTHrP assays are generally offered only in reference laboratories.

The PTH/PTHrP, whose gene (*PTHR1*) is located on chromosome 3p21.1−22, is a member of the group II subgroup of the G-protein−linked receptor (GPCR) superfamily. The PTH/PTHrP receptor has a large extracellular domain, seven transmembrane domains (e.g., heptahelical), and a cytoplasmic tail similar to the thyroid-stimulating hormone (TSH) receptor and CaSR. Other members of the GPCR group II are receptors for calcitonin, glucagon, glucagon-like peptide I, growth-hormone−releasing hormone, vasoactive intestinal polypeptide, and secretin. A second receptor that binds PTH, but not PTHrP, has been identified: PTHR2. Encoded on chromosome 2q33, PTHR2 is expressed in the brain, pancreas, testis, and placenta. The function of this receptor is yet to be fully delineated. PTHR1 appears to be the most important receptor for PTH.

The binding of PTH or PTHrP to the PTH/PTHrP receptor stimulates adenyl cyclase, raising cyclic adenosine monophosphate (cAMP) concentrations, and activating

phospholipase C. PTH stimulates osteoblasts in bone and metaphyseal or cartilaginous growth plates to foster normal bone health; in excess, PTH causes bone resorption. PTH stimulates the renal tubular cells to increase calcium reabsorption from renal tubular fluid while increasing the excretion of phosphate in the urine (e.g., increasing phosphaturia). PTH reduces expression of the sodium-dependent phosphate cotransporter [Npt2; gene name: *SLC34A1*; solute carrier family 34 (sodium phosphate), member 1; chromosome 5q35] at the renal tubular brush border that results in urinary phosphate wasting. These actions of PTH raise circulating calcium concentrations while lowering phosphate concentrations. These actions avoid the development of ectopic calcification.

Vitamin D metabolism and biology

There are two sources of vitamin D: exogenous vitamin D_2 (ergocalciferol) from the diet and endogenous vitamin D_3 (cholecalciferol) from the skin following sun exposure (Fig. 45.3). Most vitamin D comes from the skin. Vitamin D deficiency occurs in the absence of sufficient intake of vitamin D2. Over the counter, both vitamin D2 and vitamin D3 are available. Vitamin D is derived from provitamin D precursors. Provitamin D species (both D_2 and D_3) are bioconverted by ultraviolet (UV) light to form previtamin D molecules. Thermal isomerization then yields vitamin D. Provitamin D_2 is derived from the diet, whereas provitamin D_3 is produced in the skin consequent to UV light exposure. It is controversial whether vitamin D2 is less potent than vitamin D3. Vitamin D2 and vitamin D3 vary only in their R groups. As a fat-soluble vitamin, vitamin D is dependent on normal fat absorption for its absorption. Any cause of steatorrhea (e.g., bile salt deficiency or impaired secretion, or pancreatic insufficiency causing lipase and colipase deficiency) or intestinal disease (with malabsorption) can interfere with vitamin D absorption.

The primary circulating form of vitamin D, 25-hydroxyvitamin D (25-OHD), is produced in the liver via vitamin D-25-hydroxylase (gene name: *CYP27A*; chromosome 2q33-qter) in a PTH-independent manner. However, PTH does stimulate the conversion of 25-OHD to 1,25-OH_2D via 25-hydroxyvitamin D-1 alpha-hydroxylase (gene name: *CYP27B1;* chromosome: 12q1) within the renal tubular cells (Fig. 45.4). The predominant circulating form of vitamin D is 25-OHD, which has lower biological activity than 1,25-OH_2D. While clinicians might request a "vitamin D level," this is not the proper nomenclature for the measurement of 25-OHD.

Vitamin D species act by binding to the vitamin D receptor (VDR). The VDR is 427 amino acids (molecular mass ~48.3 kDa) and is encoded on chromosome 12q13.11. VDR is a member of the NR1I family whose other members include the constitutive androstane receptor and the pregnane X receptor. VDR forms heterodimers with retinoid X receptor family members. The vitamin D species with the highest affinity for the receptor is 1,25-OH_2D. As the most biologically active form of the hormone, 1,25-OH_2D functions as a transcriptional modulator. Physiologically, the active form of vitamin D also increases the absorption of calcium and phosphate from the gastrointestinal tract. 1,25-OH_2D concentrations are controlled through a short, negative feedback loop, in which 1,25-OH_2D suppresses the bioconversion of 25-OHD to the active hormone. In addition, increased phosphate

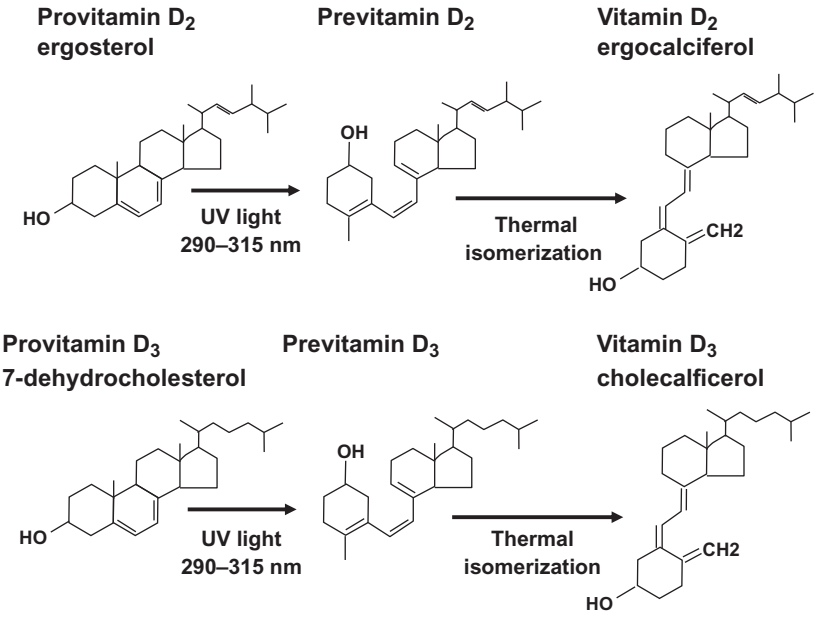


FIGURE 45.3 Provitamin D species (both D2 and D3) are acted upon by UV light to form previtamin D species. Thermal isomerization then yields vitamin D. Provitamin D2 is derived from the diet, whereas provitamin D3 is produced in the skin consequent to UV light exposure.

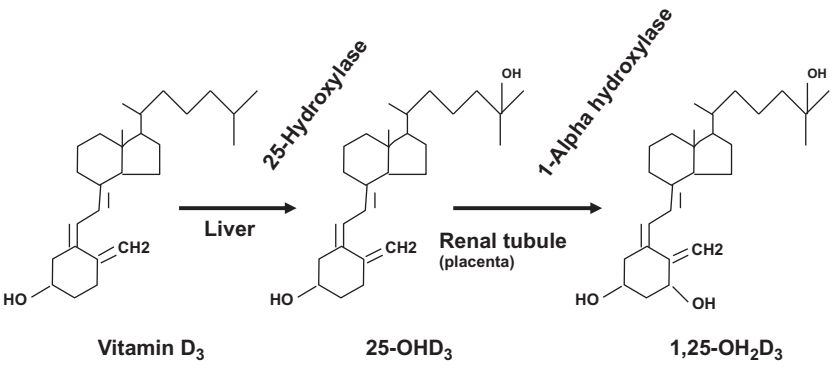


FIGURE 45.4 Vitamin D (in this case vitamin D3) is activated by a 25-hydroxylation in the liver followed by a 25-hydroxylation in the kidney. The placenta can also convert 25-OHD into 1,25-OH_2D.

concentrations and FGF-23 also decrease the conversion of 25-OHD to 1,25-OH_2D. Measurements of 1,25-OH2D are indicated only under special circumstances and such levels require expertise to interpret.

Concentrations of 1,25-OH_2D are also regulated by its conversion to 1,24,25-trihydroxy vitamin D (1,24,25-OH_3D) via the enzyme D-24-hydroxylase ("24-hydroxylase") encoded by *CYP24A1* (cytochrome P450 family 24 subfamily A member 1; chr. 20q13.2) [14]. 24-hydroxylase also converts 25-OH_2D into 24,25-dihydroxyvitamin D. 24-hydroxylase is stimulated by 1,25-OH_2D and FGF23. Loss-of-function variants in 24-hydroxylase can lead to infant hypercalcemia (sometimes referred to as "infantile idiopathic hypercalcemia").

25-OHD and 1,25-OH_2D can be measured by immunoassay or mass spectrometry. The immunoassays do not distinguish the D_2 from D_3 forms of 25-OHD and 1,25-OH_2D. However, mass spectrometry has the potential to distinguish the D_2 and D_3 forms of 25-OHD and 1,25-OH_2D.

To absorb calcium from the diet, three events occur: (1) calcium enters into the cytoplasm of the intestinal brush border cells mediated by CaT1 [a.k.a. ECaC2, an epithelial Ca^{2+} channel /transporter; gene name: *ECAC2 or TRPV6* (transient receptor potential cation channel subfamily V member 6); chromosome 7q34]; (2) calcium diffuses within these cells as mediated by calbindin Ds, which are cytosolic calcium-binding proteins [calbindin D-9K (X-chromosome) and D-28K (chromosome 8)]; and (3) from the basolateral membrane, calcium is pumped out of the cell by a CaATPase (e.g., a Na^+/Ca^{2+} exchanger). The synthesis of CaT1 is ~90% dependent on 1,25-OH_2D, whereas the synthesis of calbindin D is completely dependent on 1,25-OH_2D CaT1 and calbindin D are both rate limiting in the process of calcium absorption. On the other hand, a high calcium diet downregulates CaT1 and calbindin D by downregulating production of 1,25-OH_2D.

Similar to bone resorption, increased gastrointestinal absorption of calcium and phosphate increases both plasma calcium and phosphate concentrations. Although calcium is lost to a minor extent in the stool and urine, the primary control mechanism for total body calcium depends on the appropriate gastrointestinal tract absorption of calcium from the diet. Thus adequate calcium must be present in the diet. If excessive calcium is chronically lost in the urine (i.e., hypercalciuria), urolithiasis (kidney stones) and nephrocalcinosis (kidney calcification) can result. Either condition can lead to renal failure. In adult males, the reference interval for 24-hour urinary excretion of calcium is 25−300 mg/24 h, and in adult females, the reference interval is 20−275 mg/24 h. In adults, a random urine-calcium-to-creatinine ratio can also be used to assess the patient for hypercalciuria. A normal adult calcium-to-creatinine ratio is <0.14 (in English units); ratios >0.20 indicate hypercalciuria. Pediatric reference intervals for calcium excretion (e.g., hypercalciuria: ≥ 4 mg/kg per day) and the calcium/creatinine ratios are available in the literature [15].

Although increased phosphate absorption from the gastrointestinal tract and increased phosphate release from bone act to raise plasma phosphate concentrations, the normal phosphaturic effect of PTH on the renal tubules predominates, leading to an overall decline in plasma phosphate concentrations in response to PTH when renal function is uncompromised. Urinary excretion of phosphate is responsible for the regulation of total body phosphate.

Phosphate biology

The majority of total body phosphate (PO_4^3) resides as organic phosphate complexed with proteins, lipids, and carbohydrates. Inorganic phosphate (noncomplexed phosphate) in plasma is a very small proportion of total body phosphate.

Phosphate is plentiful in the diets of most people in industrialized countries. Gastrointestinal tract absorption of phosphate is not as highly regulated as calcium, with ~70% of dietary phosphate absorbed from the intestinal tract enterocytes (columnar epithelial cells). Excess

phosphate is then excreted in the urine. In renal failure, phosphate retention produces hyperphosphatemia. In turn, hyperphosphatemia can contribute to the development of hypocalcemia. This can be further exacerbated by decreased conversion of 25-OHD to 1,25-OH_2D, further decreasing intestinal calcium absorption. When phosphate is measured in plasma or serum, it is its elemental phosphorus content that is reported.

To summarize, the overall effects of PTH are to raise plasma calcium while lowering plasma phosphate concentrations. These actions are beneficial to avoid inappropriate biomineralization occurring in soft tissues when calcium and/or phosphate are both pathologically increased when the calcium−phosphate product may exceed $\sim 70\ mg^2/mL^2$ (i.e., calcium concentration in mg/dL is multiplied by the phosphate concentration in mg/dL). Examples of ectopic calcification include calcinosis cutis (skin calcification), tendon calcification, and blood vessel calcification. Significantly, calcified blood vessels can be visible on plain-film X-rays and CT scans (e.g., coronary calcium scans). In calcinosis cutis, firm, chip-like deposits can be palpated in the skin. Failure to maintain adequate levels of ionized calcium produces symptoms of hypocalcemia. Vice versa, excessive levels of ionized calcium produce symptoms of hypercalcemia.

Table 45.1 summarizes the major biological effects of PTH, 1,25-OH2D, and FGF23 on the body.

Bone biology

To be discussed in detail below, osteoblasts build bone, osteoclasts resorb bone, and osteocytes are osteoblasts that have encased themselves in bone creating the lacunocanalicular system. The most abundant of these cells in bone is osteocytes that comprise 90%−95% of bone cells. The next most common bone cell is the osteoblast (<5% of bone cells) and the osteoclast is the least common bone cell (1% or more of bone cells but less common than osteoblasts). Bone is in a constant state of remodeling. In adults, approximately 10% of bone turns over yearly. Initially, remodeling involves bone resorption by osteoclasts with the consequent release of calcium and phosphate and degradation of the bone protein matrix. This is followed by new bone formation (i.e., bone accretion), which is carried out by osteoblasts. Type 1 collagen is the major protein component of the bone matrix. Pyridinolines and deoxypyridinolines cross-link collagen strands.

TABLE 45.1 Causes of hypocalcemia.

1. Decreased parathyroid hormone action
2. Deficient vitamin D action
3. Extreme dietary calcium deficiency
4. The healing phase of various forms of bone disease
5. Calcium saponification (e.g., acute pancreatitis)

Osteocalcin (gene located on chromosome 1q2), also known as bone Gla protein, is the major noncollagen protein found in bone that is released from the bone matrix during bone resorption into the plasma. Synthesized by osteoblasts, osteocalcin is a cytosolic calcium-binding protein of 49 amino acids (5.7 kDa). Gla stands for "gamma-carboxy glutamic acid." Via the activity of vitamin-K−dependent carboxylases at positions 17, 21, and 24 of the protein, amino acids are modified to gamma-carboxyglutamyl residues that can bind calcium or hydroxyapatite ($Ca_3(PO_4)_2)_3.Ca(OH)_2$). Between 10% and 30% of osteocalcin produced by osteoblasts is released extracellularly. In systemic circulation, osteocalcin has a short half-life of $\sim$5 minutes. Osteocalcin concentrations can provide an index of bone turnover. Elevated osteocalcin concentrations are observed in conditions such as vitamin D deficiency states (rickets or osteomalacia), osteoporosis, hyperparathyroidism (primary or secondary, e.g., renal osteodystrophy), hyperthyroidism, or in persons with bony metastasis, acromegaly, or fractures. Osteocalcin levels decline with bisphosphonate treatment of osteoporosis or osteopenia.

Intermittent exposure to PTH normally stimulates the mesoderm-derived stromal cells of bone to develop into osteoblasts (Fig. 45.5). By entering the lacunae left by osteoclast-driven erosion of the bone microarchitectural surface, osteoblasts lay down new bone by secreting connective tissue matrix (e.g., type 1 collagen) that is ossified by calcium and phosphate deposition, resulting in the formation of hydroxyapatite. Hydroxyapatite is the hard, mineralized portion of bone. As osteoblasts are encased in this newly formed bone, the osteoblasts become osteocytes. With PTH-mediated osteoblast stimulation, bone cycling will consequently increase osteoclast activity.

Osteoclasts are ultimately derived from bone marrow hematopoietic stem cells (HSCs). HSCs give rise to common myeloid progenitor cells. Under the influence of granulocyte−monocyte colony stimulating factor, the granulocyte−monocyte progenitor arises. Subsequently, under the influence of macrophage-colony-stimulating factor (M-CSF), monocytes develop. Macrophages are formed when monocytes enter into the interstitium upon leaving the circulation.

Osteoblasts secrete M-CSF and express the receptor activator of NF-kB ligand (RANKL; see Fig. 45.5). M-CSF stimulates macrophage proliferation, while RANKL binds to the receptor activator of NF-kB on the developing osteoclast. As a member of the tumor necrosis factor superfamily, RANKL is also recognized as an "osteoclast

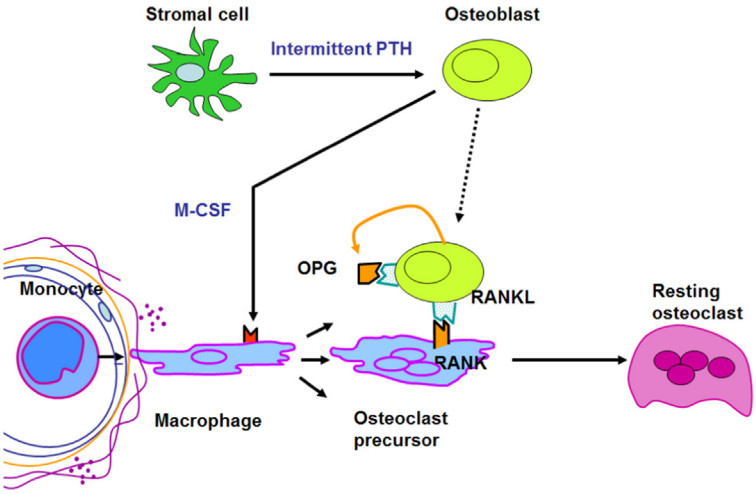


FIGURE 45.5 Under the influence of intermittent PTH stimulation, stromal cells develop into osteoblasts. Via their secretion of macrophage-colony-stimulating factor, macrophages develop into osteoclast precursors. Through the expression of RANKL binding to RANK, the osteoblast precursors develop into resting osteoclasts. Opposing the interaction of RANKL and RANK, osteoblasts secrete osteoprotegerin, which serves as a soluble decoy receptor for RANKL. Therefore the number of "free" RANKL molecules determine the degree to which the osteoblast stimulates the osteoclast.

differentiation factor" or osteoprotegerin (OPG) ligand. Therefore cell-surface proteins on the osteoblast stimulate osteoclast development, thereby fostering bone remodeling. This firmly links osteoblast and osteoclast activity.

RANKL stimulates the differentiation of osteoclasts. This development is also enhanced by intermittent exposure to PTH. On the other hand, osteoblasts secrete OPG, which binds to RANKL and inhibits its action by functioning as a soluble decoy receptor for RANKL. A case study described a young man presenting with osteoporosis who produced an autoantibody against OPG; this process theoretically enhanced osteoclast proliferation, causing his osteoporosis [16]. Deficiency of OPG is one cause of juvenile Paget disease of bone [17]. The expression of RANKL by osteoblasts is increased by cytokines such as IL-1, IL-6, and IL-11, explaining the common finding of osteopenia in chronic inflammatory conditions (e.g., rheumatoid arthritis).

Attachment of the osteoclast to bone occurs via osteoclastic integrins ($alpha_v beta_3$) located on foot-like, actin-containing podosomes that bind to osteopontin and vitronectin on the bone's surface to produce a tight seal (Fig. 45.6). The space between the osteoclast and the bone is termed the subosteoclastic compartment, which is also known as the Howship lacunae. Within the osteoclast, carbon dioxide and water react to form hydrogen (H^+) and HCO_3^- due to carbonic anhydrase II catalysis. At the apical side of the cell, chloride (Cl^-) and HCO_3^- are exchanged via the HCO_3^- −chloride exchanger. At the ruffled border of the osteoclast, a chloride channel allows entry of chloride into the subosteoclastic compartment accompanying H^+ transported into the subosteoclastic compartment by a H^+-ATPase pump (a vacuolar-type ATPase).

Secreted cathepsin k (a cysteine protease; a member of the peptidase C1 protein family; 329 amino acids; ~37 kDa; gene on chromosome 1q21.3) and matrix metalloprotease-9 (707 amino acids; ~78.5 kDa; gene on chromosome 20q13.12) degrade the connective tissue matrix, whereas hydrochloric acid creates an acidic (pH of 4.5) environment to afford solubilization of the mineral content of the bone. Degraded bone components exit the subosteoclastic compartment via transcytosis to the interstitium. The released calcium and phosphate increase calcium and phosphate concentrations in plasma.

Whereas normal concentrations and a pulsatile release of PTH are vital to maintain healthy bone remodeling, persistent excess PTH increases RANKL and decreases OPG production by osteoblasts overall, leading to excessive osteoclast formation. When osteoclast activity exceeds osteoblast activity, bone resorption occurs, resulting in a decline in the mineral content of bone with loss of bone strength. A decline in bone strength predisposes to osteopenic fractures.

Stimulated osteoblasts secrete increased concentrations of bone alkaline phosphatase (BAP). BAP can be measured by immunoassay. Other sources of alkaline phosphatase include [1] the biliary tract with elevations indicative of biliary tract disease and [2] the placenta where alkaline phosphatase is normally elevated in the third trimester of pregnancy. While alkaline phosphatase isoenzymes can be measured by electrophoresis (or the even older inhibition techniques), there are usually better markers of bone or biliary tract disease that can be ordered instead of ordering alkaline phosphatase isoenzyme measurements. Increased BAP is found in childhood and is associated with growth. Further, BAP raises the following fractures in all age groups. Transient hyperphosphatasemia of infancy and childhood describes a condition of unknown etiology, where total alkaline phosphatase activity measurements are inexplicably elevated at least four to five times greater above the upper limit of the reference interval for children (sometimes >10 times the upper limit of the reference interval) [18].

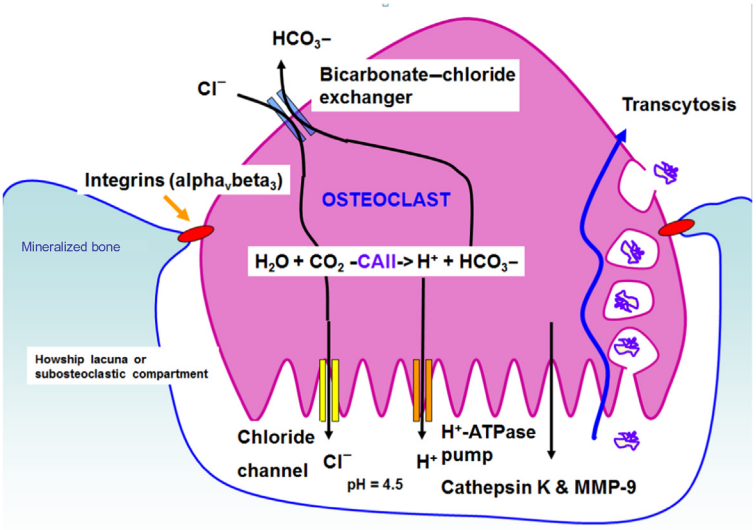


FIGURE 45.6 Once osteoclasts bind to bone, hydrochloric acid, cathepsin K, and matrix metalloprotease-9 enter the subosteoclastic compartment to resorb bone. The released materials transcytose across the osteoclast.

Increased BAP or osteocalcin concentrations signal an increased rate of bone formation although bone loss can exceed bone formation leading to an overall decline in bone mineral density (BMD).

In adults, maintenance of normal bone mass requires proper nutrition in the form of adequate dietary vitamin D and calcium, weight-bearing exercise, sex steroids in pubertal and postpubertal individuals, and growth hormone. Decreased bone density is described as osteopenia. By itself, osteopenia is not a disease. Osteopenia can result in impaired bone strength and a propensity to fractures. Radiologically, osteopenia can be observed with the loss of ~50% of BMD. A more sensitive measure of bone mineral content is dual-energy X-ray absorptimetry (DEXA scanning). DEXA scans are commonly used to measure BMD in perimenopausal and menopausal women.

Disorders that involve the bone connective tissue matrix can also interfere with bone formation, turnover, and strength. Increased bone turnover with type 1 collagen breakdown can be assessed by detecting increased concentrations in the urine and/or serum of pyridinolines, deoxypyridinolines, N-telopeptides, and C-telopeptides [19]. C-telopeptides are measured in serum, N-telopeptides can be measured in serum or urine, and pyridinolines and deoxypyridinolines are usually measured in timed urine collections. It remains controversial whether serial measurements of bone markers improve the diagnosis or management of osteoporosis [20].

The diagnosis of osteoporosis is made clinically and radiologically, where the patient's BMD by DEXA scan displays a T-score of more than 2.5 standard deviations below the mean. In the absence of fractures, routine laboratory testing is normal (e.g., calcium, phosphate, alkaline phosphatase, and creatinine). Causes of secondary osteopenia such as primary hyperparathyroidism, secondary hyperparathyroidism of renal failure, and osteomalacia must be excluded clinically and through laboratory testing.

Hypocalcemia

Electrophysiologically, hypocalcemia lowers the threshold for depolarization of the sodium channel. This accounts for hyperreflexia and tetany in patients with significant degrees of hypocalcemia. The signs of hypocalcemia also include muscle pain or cramps, carpopedal spasm, paresthesias, seizures, and cardiac arrhythmias evidenced in a prolonged QT interval. The Chvostek sign is demonstrated by lightly tapping the face between the ear and temporomandibular joint eliciting an ipsilateral (e.g., the same side) facial contraction. The Trousseau sign of hypocalcemia is elicited by occluding the brachial artery with a blood pressure cuff, leading to spasm of the hand and forearm within 3 minutes. Chronic hypocalcemia can be associated with basal ganglia and corneal calcifications. There are numerous causes for hypocalcemia (Table 45.1).

Decreased parathyroid hormone action

With decreased PTH action, hypocalcemia ultimately results from a failure to absorb sufficient calcium from the gastrointestinal tract, failure to release calcium from bone, and calcium wasting in the urine. The most common cause of decreased PTH action is hypoparathyroidism (e.g., decreased PTH secretion by the parathyroid gland). In the absence of PTH's phosphaturic effect, PTH deficiency causes phosphate retention and hyperphosphatemia.

Hypoparathyroidism

The laboratory diagnosis of hypoparathyroidism is based on the recognition of a low PTH concentration in the setting of hypocalcemia. Hyperphosphatemia supports the diagnosis of hypoparathyroidism. Intact PTH and whole PTH assays are equally effective at identifying low PTH concentrations in plasma or serum. Whereas some research papers report nearly identical reference intervals for intact PTH and whole PTH assays, most research papers report much different reference intervals with the whole PTH reference interval being approximately 40%–50% lower than the intact PTH reference interval.

Hypoparathyroidism has many etiologies including autoimmune parathyroid destruction, parathyroidectomy, irradiation, hypoparathyroidism associated with sensorineural and renal dysplasia, hypoparathyroidism associated with retardation and a dysmorphic appearance [hypoparathyroidism-retardation dysmorphism (HRD)], hypoparathyroidism associated with various other genetic conditions (e.g., DiGeorge syndrome, ring chromosome 16 and ring chromosome 18), and idiopathic hypoparathyroidism. HRD results from a mutation in tubulin cofactor E (del52–55). There are a large number of genetic conditions associated with hypoparathyroidism (Table 45.2).

Chronic hypomagnesemia can impair both PTH release and tissue response to PTH. Familial hypoparathyroidism can display many forms of Mendelian inheritance, such as autosomal dominant (AD), autosomal recessive (AR), and X-linked varieties. AD hypoparathyroidism with normal to increased urinary calcium excretion can result from gain-of-function variants within the *CaSR* gene, leading to inappropriately low PTH secretion due to parathyroid hypersensitivity to ionized calcium concentrations. This is sometimes referred to as "familial hypercalciuric hypoparathyroidism (FHH) type 1." The variants that cause FHH type 1 include missense mutations of the extracellular domain, first extracellular loop, and the fifth and sixth transmembrane domains, or an in-frame 181-bp deletion of the CaSR cytoplasmic tail. FHH types 2 and 3 do not result from CaSR gain-of-function mutations. FHH type 2 is linked to a locus on chromosome 19p13, whereas FHH type 3 is linked to a locus on chromosome 19q13. PTH gene mutations have been described as causes of AD (a signal peptide mutation) and AR (an exon–intron boundary mutation) familial isolated hypoparathyroidism.

DiGeorge syndrome results from a malformation of the third and fourth pharyngeal pouches. In addition to hypoparathyroidism, abnormalities of the aortic arch, thymus, and facial structures (e.g., low-set ears, a beak-shaped mouth) may be observed. Thymic hypoplasia or aplasia leads to mild to severe T-cell immune deficiencies expressed in increased frequencies of viral and fungal infections. A microdeletion of the long arm of chromosome 22 (i.e., 22q11) is the cause of DiGeorge syndrome. If the deletion cannot be detected by karyotypic analysis, the deletion can be evaluated by fluorescence-labeled in situ hybridization. New technologies under study to detect deletions include comparative genomic hybridization [21].

Autoimmune hypoparathyroidism can exist as an isolated condition or can be part of autoimmune polyglandular syndrome type I (APS type I) [22]. Autoantibodies reactive with the parathyroid gland are rare at best [23]. Autoantibodies against the CaSR that stimulate or block the receptor are well described but likely do not identify parathyroid gland destruction [24]. This is discussed again later in this section. APS type I is the concurrence of hypoparathyroidism, mucocutaneous candidiasis, and Addison disease or adrenal autoantibodies. Two of these three conditions must be present to diagnose APS type I. This early-onset AR disorder is caused by mutations in the autoimmune regulator (*AIRE*) gene.

TABLE 45.2 Genetic, syndromic, or familial disorders with hypoparathyroidism.

Hypoparathyroidism associated with sensorineural and renal dysplasia
Hypoparathyroidism associated with retardation and dysmorphic appearance
Hypoparathyroidism associated with ring chromosome 16
Hypoparathyroidism associated with ring chromosome 18
DiGeorge syndrome
Autoimmune polyglandular syndrome type 1
Autosomal dominant hypoparathyroidism, type 1 (CaSR, gain-of-function)
Autosomal dominant hypoparathyroidism, type 2 (chromosome 19p13)
Autosomal dominant hypoparathyroidism, type 3 (chromosome 19q13)
Autosomal recessive hypoparathyroidism
X-linked hypoparathyroidism

CaSR, Calcium-sensing receptor.

The AIRE protein is a transcription factor. AIRE is a 545 amino acid protein with a molecular mass of approximately 58 kDa that is encoded on chromosome 21q22.3. AIRE elicits its function in a dimeric or tetrameric form. The normal action of AIRE relates to the expression of self-peptides in the thymus. At the present time, APS type I is the only known single-gene, autosomal recessive autoimmune disease in humans. Other associated diseases in APS type I patients include gonaditis (producing primary gonadal failure in women) and autoimmune hepatitis. Less commonly associated conditions are type 1 diabetes mellitus, autoimmune thyroid disease, vitiligo, alopecia, fat malabsorption, IgA deficiency, pernicious anemia, red cell aplasia, and progressive myopathy. APS type I has also been named "autoimmune polyendocrinopathy−candidiasis−ectodermal dystrophy syndrome," because associated disorders include nail dystrophy and dental enamel hypoplasia.

Hypocalcemia has been reported infrequently in another immunopolyglandular disorder: the immunodysregulation, polyendocrinopathy, and enteropathy, X-linked (IPEX) syndrome [25]. Disorders in this syndrome include early-onset type 1 diabetes mellitus, severe enteropathy with diarrhea, eczema, autoimmune hemolytic anemia, thrombocytopenia, and thyroiditis or hyperthyroidism. IPEX is usually lethal early in life. Of two IPEX patients reported with hypocalcemia, in one case, the PTH was normal, whereas in the other case, the PTH was increased, suggesting possible PTH resistance. Variants in the transcription factor protein forkhead box P3 (FOXP3) cause IPEX. FOXP3 (chromosome Xp11.23, encoding scurfin) is involved in the development of regulatory CD4 + CD25 + T cells. Such T regulatory cells (Tregs) can form in the thymus ("natural" Tregs) or they can be induced in the periphery ("induced" Tregs).

Currently, there are no readily available clinical tests for parathyroid gland autoantibodies, although target autoantigens have been demonstrated. In the 1960s, parathyroid autoantibodies were detected by indirect immunofluorescence using parathyroid gland as a substrate. However, these findings have been difficult to confirm in subsequent studies. Autoantibodies to NACHT (neuronal apoptosis inhibitory protein, MHC class II transcription activator, HET-E, and TP1) leucine-rich-repeat protein 5 were described in ∼50% of hypoparathyroid patients with APS type 1. NACHT refers to a 300−400 amino acid domain with predicted nucleoside triphosphatase activity [26]. Using Western blotting, autoantibodies to the CaSR protein have been described [27]. Independent of autoimmune parathyroiditis, autoantibodies against the CaSR have been identified that can block this receptor (resulting in hyperparathyroidism) or stimulate the receptor (causing hypoparathyroidism). Because of the current limitations in the availability of autoantibody testing for hypoparathyroidism, to make the diagnosis of autoimmune hypoparathyroidism, the clinician must rely on the presence of associated autoantibodies (e.g., adrenal autoantibodies or interferon-ω autoantibodies in APS I) or autoimmune conditions such as Addison disease to implicate autoimmunity as a cause of hypoparathyroidism [28].

Transient hypoparathyroidism may occur in the settings of prematurity, infants of diabetic mothers, neonatal illness, and children or adults with severe illness or stress such as recent surgery or sepsis. After recovery from the inciting illness, normal parathyroid function and normocalcemia usually returns spontaneously.

Pseudohypoparathyroidism

When hypocalcemia and hyperphosphatemia are present in a patient whose PTH concentrations are above the reference interval, PTH resistance must be considered. Resistance to PTH is termed pseudohypoparathyroidism (PHP). PHP is currently described as decreased activation of the G-stimulatory protein α subunit/cAMP/protein kinase A (Gsα/cAMP/PKA) signaling pathway by PTH (and other hormones) that bind to Gsα-coupled receptors [29].

Historically, PHP has been classified as type 1 or type 2. In type 1 PHP, the defect is in the generation of cAMP, which is a secondary messenger for the PTH/PTHrP receptor. Traditionally, this condition was identified by the inability of an exogenous PTH injection to elicit either increased cAMP concentrations in the urine or concurrent phosphaturia. In PHP type 1a (PHP1A), there is a maternal loss-of-function mutation in *GNAS*, the gene that encodes Gsα. The clinical phenotype of PHP1A includes resistance to PTH, TSH, and gonadotropins, brachydactyly (especially, shortening of the fourth and/or fifth metacarpals), heart-shaped facial features, progressively decreasing growth velocity causing adult short stature, early-onset obesity, neurocognitive impairment, and cerebral calcifications [30]. These features are described as "Albright hereditary osteodystrophy" (AHO).

Isolated resistance to PTH, absence of clinical features of AHO, and normal Gsα activity initially described PHP1B. For example, persons with PHP1B do not display adult short stature or resistance to gonadotropins. However, upon further study, some persons with PHP1B can share various features that overlap with PHPA1, such a brachydactyly, TSH resistance, and cerebral calcifications. Whether a third variety of PHP exists (PHP1C) is controversial and the 2018 Consensus Statement declined to address PHP1C because of a lack of firm data about this entity [30].

In the type 2 variety of PHP (PHP2), the defect is in the cell's response to cAMP. Following injected PTH, increased cAMP excretion in the absence of increased

urinary phosphate excretion was consistent with cAMP resistance. If increased cAMP and phosphate were both recognized after exogenous PTH injection, the possibility of a defective PTH molecule was considered although such inborn errors are very rare. The molecular cause or causes of PHP2 are unknown.

Physical examination and developmental features of AHO, in the absence of disordered calcium/phosphate metabolism, identify pseudopseudohypoparathyroidism (PPHP). Reduced Gsα activity in the absence of PTH resistance has been demonstrated in PPHP. In PPHP, the *GNAS* loss-of-function mutation is inherited from the father (whereas *GNAS* in PHPH1A loss-of-function mutation is inherited from the mother). Families with both PHP and PPHP have been reported but there are no molecular studies of these families [31].

Other defects in the Gsα/cAMP/PKA signaling pathway have been reported. Progressive osseous heteroplasia results from *GNAS* mutations that cause ossification of skin and muscle. However, tissue responses to PTH and TSH are normal. Another rare syndrome, acrodysostosis, is identified clinically by facial dysostosis, nasal hypoplasia, and severe brachydactyly together with TSH and PTH resistance. Acrodysostosis type 1 (ACRDYS1) results from mutations in the *PRKAR1A* (protein kinase cAMP-dependent type I regulatory subunit alpha; chr. 17q24.2) gene, causing decreased PKA activity [32]. Normally, cAMP binds to the regulatory subunits of PKA causing the inactive PKA (composed of two regulatory subunits and two catalytic subunits) to dissociate into two dimers: a dimer of regulatory subunits and a dimer of catalytic subunits. Three different catalytic subunits and four different regulatory subunits of PKA have been described. Active PKA (i.e., the "free" catalytic dimers) phosphorylates various proteins such as phosphodiesterases (e.g., *PDE3A* and *PDE4D*) and cAMP-responsive binding elements. Phosphodiesterases normally lead to degradation of cAMP. Acrodysostosis type 2 (ACRDYS2) results from gain-of-function mutations in *PDE4D* [33] that could result in reduced cAMP levels because of accelerated cAMP clearance. ACRDYS1 and ACRDYS2 are clinically indistinguishable. Because ACRDYS1 and ACRDYS2 share many physical characteristics with PHP such as brachydactyly, obesity, and short stature, ACRDYS1 and ACRDYS2 can be considered phenocopies of PHP.

Individuals who display bone sensitivity to PTH but renal resistance to PTH are also described as having pseudohypohyperparathyroidism. Because of renal resistance, systemic calcium concentrations decline and phosphate concentrations rise, eliciting increased PTH secretion. However, because of bone sensitivity to PTH, bone resorption occurs similar to hyperparathyroid states. There are no reports of molecular studies of this rare disorder [34].

Deficient vitamin D action

In the setting of deficient vitamin D action, there is inadequate absorption of calcium from the diet. Deficient vitamin D action can occur because of dietary vitamin D deficiency, vitamin D malabsorption, and deficient transformation of vitamin D to its most active form (i.e., 1,25-OH2D), VDR defects. Inadequate absorption of calcium causes hypocalcemia. The parathyroid gland's response to hypocalcemia (via the CaSR) is the release of PTH. This stimulates increased conversion of 25-OHD to 1,25-OH_2D, resulting in increased calcium absorption and bone resorption. These actions are usually sufficient to raise extracellular calcium concentrations back into the normal range. In the background of secondary hyperparathyroidism, however, there is increased phosphaturia, resulting in hypophosphatemia leading to inadequate bone matrix mineralization and rickets (in children) or osteomalacia (in adults). In rickets or osteomalacia, total calcium is usually in the lower reference interval.

Increased osteoblast activity, as a consequence of increased PTH concentrations, causes an increase in the bone isoenzyme concentration of alkaline phosphatase. BAP is best measured by immunoassay. Therefore the classic findings of rickets after PTH compensation has occurred (e.g., secondary hyperparathyroidism is present) are a low-normal calcium, low phosphate, increased PTH, and increased total alkaline phosphatase. Inadequate extracellular concentrations of calcium and/or phosphate interfere with normal bone formation in growing children, producing rickets. Rickets is characterized by metaphyseal widening, irregular epiphyseal surfaces, bowing of weight-bearing bones (e.g., the legs), osteopenia, increased prominence of the costochondral junctions (e.g., "rachitic" rosary), and an indentation of the lower rib cage due to bone softening and the pull of the diaphragm on the rib cage (Harrison's groove, a.k.a. Harrison's sulcus).

Vitamin D-dependent rickets result from either defective conversion of 25-OHD to 1,25-OH_2D because of an autosomal recessively inherited error in the renal tubular 1 alpha-hydroxylase enzyme or end-organ resistance to the effects of 1,25-OH_2D. Whereas vitamin D-deficient rickets can be ultimately cured by ensuring that the recommended daily allowance of vitamin D (400 IU/day) is taken in the diet, vitamin D-dependent rickets requires pharmacologic doses of vitamin D (e.g., 2000 IU/day) for correction of rickets or osteomalacia (to be discussed later). Such large doses will overcome the inborn metabolic block in the conversion of 25-OHD to 1,25-OH_2D or the receptor resistance to 1,25-OH_2D. The two forms of vitamin D-dependent rickets can be separated by measuring 1,25-OH_2D concentrations. With decreased conversion of 25-OHD to 1,25-OH_2D, the 1,25-OH_2D

concentration is inappropriately low. However, in the background of 1,25-OH_2D resistance, concentrations are above the reference interval.

Hypophosphatemic rickets arise from a variety of renal tubular disorders where there is excessive phosphaturia, producing hypophosphatemia and insufficient phosphate to ossify bone. Hypophosphatemic (vitamin D-resistant) rickets is most commonly inherited as an X-linked dominant disorder caused by mutations in *PHEX*, and less commonly as an AD disorder caused by *FGF23* mutations, an X-linked recessive disorder (XLR) caused by *CLCN5* mutations, or as an AR disorder caused by mutations in *GNAS*, dentin matrix acidic phosphoprotein (*DMP1*), or *SLC34A3*.

In X-linked dominant familial hypophosphatemic rickets (XHR), a variant in the cell-membrane Zn-metalloendopeptidase *PHEX* (phosphate regulating gene with homologies to endopeptidases on the X chromosome; 749 amino acids; molecular mass ~86.5 kDa; gene location: chromosome Xp22.11) is the etiology. *PHEX* mutations elevate FGF23, which produces hyperphosphaturia. XHR can also be referred to as XHR. Normal sources of PHEX are osteoblasts, muscle, lung, liver, and the gonads. PHEX structurally resembles the neutral endopeptidases endothelin-converting enzyme-1 and Kell antigen. A normal function of PHEX is to inhibit cleavage of matrix extracellular phosphoglycoprotein (MEPE). When MEPE is cleaved, an acidic serine aspartate-rich MEPE-associated motif (ASARM) is produced that inhibits NPT2, similar to the action of FGF23. With less inhibition of MEPE cleavage, higher ASARM levels are achieved, causing more phosphaturia. Because elevated FGF23 impairs the conversion of 25-OHD into 1,25-OH2D, XHR produces a state of vitamin D deficiency.

X-linked recessive hypophosphatemic rickets results from *CLCN5* loss-of-function mutations. *CLCN5* (chloride channel, voltage-sensitive 5; 746 amino acids, molecular mass ~83.1 kDa; chromosome Xp11.23) encodes a voltage-gated proximal renal tubular chloride channel that is expressed in intracellular endosomes. Other conditions caused by *CLCN5* loss-of-function mutations include X-linked nephrocalcinosis, nephrolithiasis, renal failure, and Dent disease (a type of chronic kidney disorder in males).

The autosomal dominant form of hypophosphatemic rickets is caused by *FGF23* mutations. FGF23 is 251 amino acids (molecular mass ~30.0 kDa) and is encoded on chromosome 12p13.32. The mutated FGF23 is resistant to degradation and FGF23 concentrations rise causing hyperphosphaturia and hypophosphatemia. Similar to XLR, there is impaired conversion of 25-OHD into 1,25-OH2D because of the elevated FGF23 concentrations. The specific enzymes degrading FGF23 are unknown at the present time.

In general, FGFs affect cell survival and cell division involving the development and differentiation of tissues, healing, and carcinogenesis. The FGF family includes more than 20 members. FGFs bind to FGF receptors (FGFR), including FGFR1, FGFR2, FGFR3, and FGFR4. FGF19, FGF21, and FGF23 (the endocrine FGFs) represent one of the seven subfamilies of FGFs based on amino acid composition, sequence similarity, and implied evolutionary similarities. The hypophosphatemic action of FGF23 is mediated by its binding to FGFR1 with interaction involving the protein Klotho (to be discussed later). Of the AR forms of hypophosphatemic rickets (ARHR), *GNAS* gain-of-function mutations increase the effects of FGF23. In addition to ARHR, gain-of-function GNAS mutations can cause fibrous dysplasia of bone and McCune-Albright syndrome.

DMP1 loss-of-function mutations increase FGF23 concentrations [35]. DMP1 is a noncollagenous tooth and bone protein and belongs to a class of proteins termed small integrin-binding ligand, N-linked glycoproteins that include osteopontin and bone sialoprotein. DMP1 is 513 amino acids (molecular mass ~55.8 kDa) and is encoded on chromosome 4q22.1. *SLC34A3* [solute carrier family 34 (type II sodium/phosphate cotransporter), member 3; a.k.a. sodium-dependent phosphate cotransporter type IIC] loss-of-function mutations impair renal tubular phosphate reabsorption, causing hyperphosphaturia. SLC34A3 is 599 amino acids (molecular mass ~63.6 kDa) encoded on chromosome 9q34.3. *SLC34A3* mutations also cause hypercalciuria.

Deficient extracellular phosphate concentrations cause hypophosphatemic rickets. This disorder cannot be healed through the administration of vitamin D alone. Because of excessive phosphaturia, patients with this disorder must take supplemental oral phosphate up to six times per day. Unfortunately, such medications are of low palatability. Nevertheless, if sufficient oral phosphate is ingested, along with a modest dose of vitamin D to support phosphate absorption from the gastrointestinal tract, rickets can be substantially treated when plasma phosphate concentrations exceed ~3 mg/dL between the doses of oral phosphate.

Any renal tubular disorder that results in hyperphosphaturia can also produce rickets. One such example is Fanconi syndrome. In Fanconi syndrome affecting the renal tubules, in addition to phosphate loss, electrolytes, glucose, amino acids, and HCO_3^- are also lost in the urine. Urinary glucose loss in the absence of hyperglycemia is termed "renal glycosuria," which results from *SGLT2* mutations [SLC5A2; solute carrier family 5 (sodium/glucose cotransporter), member 2 encoded on chromosome 16p11.2]. HCO_3^- loss produces renal tubular acidosis, in which there is a normal anion gap and hyperchloremia.

Rickets can develop in the cases of obstructive biliary tract disease for several reasons. With hepatocyte dysfunction, there may be decreased conversion of vitamin D

to 25-OHD. However, the major reason for vitamin D deficiency in biliary tract obstruction (with reduced bile delivery to the intestine) is fat malabsorption that leads to malabsorption of fat-soluble vitamins. Persons with biliary tract disease may also suffer from deficiencies in vitamin A, E, and K. Although controversial, drugs that stimulate the P450 enzyme system appear to be associated with vitamin D deficiency, possibly by increasing the metabolism of vitamin D [36]. Classic drugs to cause "toxic hepatic hydroxylation" include phenytoin and phenobarbital.

Vitamin D deficiency in adults may lead to osteomalacia, which can cause osteopenia, bone pain, bone tenderness, proximal muscle weakness, bone deformations, and fractures. However, compared with children with rickets, osteomalacia causes less severe bony deformities. While there is tremendous interest in the nonskeletal effects of vitamin D, there is no firm evidence that vitamin D, as a treatment, is useful for any disorders other than states of vitamin D deficiency affecting the skeleton [37]. Until recently, dietary deficiency of vitamin D was thought to be very uncommon in industrialized countries. However, a rising frequency of vitamin D deficiency is being observed. For example, in the absence of vitamin D supplementation, exclusively breast-fed infants are at risk of nutritional rickets. Low concentrations of 25-OHD are detected in conditions of dietary vitamin D deficiency. Adjusting the reference interval for 25-OHD according to the PTH level may be necessary.

When selecting an assay to measure 25-OHD, the assay should measure total 25-OHD (e.g., 25-OHD2 plus 25-OHD3) or the assay can measure each analyte uniquely. However, there are no data to suggest that separate monitoring of 25-OHD2 and 25-OHD3 is helpful in the diagnosis or follow-up of vitamin D deficiency. Analytic problems with 25-OHD measurements persist. Bias exists between assays, in part, because of a lack of standardization [38]. Various laboratories can report twofold differences in 25-OHD concentrations for the same sample. A 2013 paper from the National Institute of Standards and Technology [39] recommended using Standard Reference Material 972a and Standard Reference Material 2972 for vitamin D measurements. These materials have certified concentration values for vitamin D metabolites. Some assays are insensitive and are poorly able to detect 20% changes in 25-OHD concentrations in serially measured samples. Assays may not be equipotent in measuring 25-OHD_2 versus 25-OHD_3. Many laboratories have turned to mass spectrometry to measure 25-OHD_2 and 25-OHD_3.

What constitutes a sufficient 25-OHD level remains controversial [40–43]. Whereas reference intervals for 25-OHD are reported in the teens or lower (e.g., 13–67 ng/mL for children and 9–54 ng/mL in adults), maximal absorption of dietary calcium is achieved with 25-OHD levels near 30 ng/mL. This suggests that a healthy 25-OHD level should be at or above 30 ng/mL. Some aggressive clinicians are routinely prescribing up to 2000 IU of vitamin D per day to achieve higher 25-OHD levels in the hope that this will lead to improved bone strength, resistance to fractures, and less osteoporosis. On the other hand, if hypercalcemia were to develop as a consequence of high-dose vitamin D therapy, significant morbidity could ensue.

The Institute of Medicine [44] concluded in 2011 that the minimum 25-OHD concentration sufficient to meet the needs of 97.5% of the population was 20 ng/mL (50 nmol/L). Clinical practice guidelines from the Endocrine Society describe 25-OHD concentrations of <20 ng/mL as deficient, whereas concentrations between 21 and 29 ng/mL are insufficient [45]. Mayo Clinical Laboratories defines severe vitamin D deficiency as 25-OHD concentrations as <10 ng/mL and mild to moderate vitamin D deficiency as 10–19 ng/mL [46]. Quest Laboratories provides a 25-OHD reference interval of 30–100 ng/mL for chromatography/mass spectrometry [47]. LabCorp uses the same reference interval [48]. Finally, ARUP Laboratory describes vitamin D deficiency as a 25-OHD concentration of <20 ng/dL and insufficiency is 20–29 ng/mL [49].

The condition "tumor-induced" osteomalacia (a.k.a. oncogenic osteomalacia) results from tumors that produce FGF23 that induce hypophosphatemia because of hyperphosphaturia. Causes of this condition include hemangiopericytoma, phosphaturic mesenchymal tumor of the mixed connective tissue type, osteoblastoma-like tumors, and fibroma-like tumors (ossifying and nonossifying). In such cases, the identification of an elevated FGF23 concentration is diagnostically helpful.

Other causes of hypocalcemia

In order for vitamin D to serve its purpose, there must be sufficient dietary calcium for absorption. Thus a rare cause of hypocalcemia is inadequate calcium intake. Another uncommon cause of hypocalcemia is hemorrhagic and edematous pancreatitis. In cases of pancreatitis, lipase is released extracellularly due to acinar cell necrosis. Increased lipase activity can lead to increased breakdown of fats to free fatty acids and glycerol. Increased concentrations of free fatty acids can bind plasma calcium, reducing circulating total calcium concentrations. This extraction of calcium from the bloodstream by the tissues (e.g., saponification) causes hypocalcemia. Another rare cause of hypocalcemia is observed following the treatment of hyperparathyroidism, hyperthyroidism, or hematologic malignancies, where healing bone consumes increased amounts of calcium for

new bone formation. This can be referred to as "hungry bone syndrome."

Certain drugs can cause hypocalcemia. Bisphosphonates used in the treatment of osteoporosis reduce bone turnover by inhibiting osteoclasts. With a decline in calcium concentrations, PTH release is triggered, subsequently increasing the conversion of 25-OHD into 1,25-OH_2D, increasing urinary calcium reabsorption, and increasing urinary phosphate excretion. The newest bisphosphonates are taken monthly; once-yearly intravenous (IV) bisphosphonates were approved in 2007. An uncommon but serious complication of bisphosphonates is osteonecrosis of the jaw. Bisphosphates can also cause adynamic bone disease.

Adynamic bone disease is a type of renal osteodystrophy with decreased numbers of osteoblasts and osteoclasts, no increased osteoid, and very low bone turnover [50].

During medical procedures such as apheresis (whether for collection or therapy), calcium concentrations can decrease transiently [51]. For instance, during a plateletpheresis procedure, there may be a 6%−25% reduction in ionized calcium. Dilution, redistribution, metabolism, and excretion of infused citrate used as the anticoagulant during the procedure are important factors protecting against immediate extreme levels of hypocalcemia. Supplemental calcium may also be given during these procedures to mitigate the physiologic effects of hypocalcemia. PTH rises quickly in the first 5−15 minutes then more slowly during the next 30−90 minutes of these procedures. Serum and urine citrate levels ordinarily return to baseline within 4 hours after infusion ceases.

In the treatment of secondary hyperparathyroidism, a family of drugs have been developed that bind to the CaSR to reduce PTH concentrations. These CaSR agonists suppress PTH secretion (which could uncommonly induce hypocalcemia). The first oral drug in this family was cinacalcet hydrochloride (Sensipar in the United States, Australia, New Zealand, and Canada and Mimpara in Europe). By reducing PTH secretion, cinacalcet hydrochloride lowers calcium, phosphate, and the calcium−phosphate product (also reducing ectopic calcification).

Clinical approach to hypocalcemia

If a patient's total plasma calcium is decreased, plasma albumin or, preferably, ionized calcium should be measured to confirm that the hypocalcemia is not due to hypoproteinemia (e.g., hypoalbuminemia). Next, phosphate, magnesium, and PTH should be measured and renal function assessed (e.g., creatinine, BUN, and urinalysis), because renal dysfunction can have a profound effect on calcium and bone metabolism.

Clinically, the causes of hypocalcemia can initially be divided into those conditions associated with increased PTH and those conditions with low PTH (Fig. 45.7). Hypocalcemia and hyperphosphatemia are most consistent with functional hypoparathyroidism. If the PTH is indeed low, hypoparathyroidism is diagnosed. Chronic hypomagnesemia must also be excluded. Hypomagnesemia is usually associated with normal or, in 20% of cases, low phosphate concentrations, although chronic hypomagnesemia does decrease both PTH release and PTH response. If PTH is paradoxically normal or high in the face of hypocalcemia and hyperphosphatemia, PTH resistance is identified and PHP and its varieties should be considered.

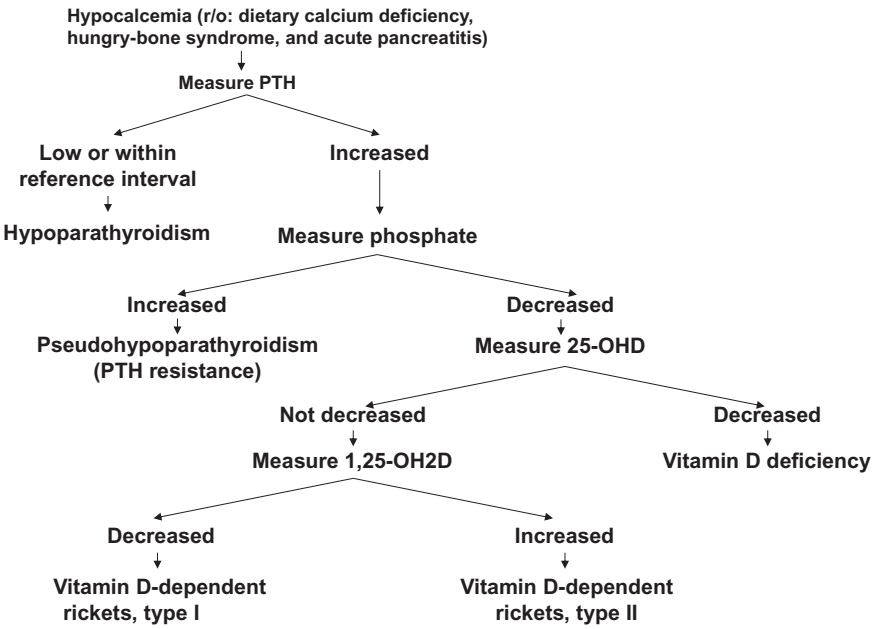


FIGURE 45.7 Diagrammed is an approach to the evaluation of hypocalcemia. See the text for details.

Low or low normal calcium, hypophosphatemia, and an increased PTH level are consistent with rickets in children or osteomalacia in adults. A depressed 25-OHD level and a history of poor vitamin D intake or lack of exposure to sunlight establish the diagnosis of vitamin D deficiency. Cure of the rickets or osteomalacia with 400 IU of vitamin D daily supports the diagnosis to vitamin D deficiency. However, physicians will usually give higher doses of vitamin D or vitamin D analogs for a short period of time to enhance initial bone healing in vitamin D deficiency. When routine doses of vitamin D do not heal a child's rickets, vitamin D-dependent rickets should be considered. Radiologic healing with ~2000 IU of vitamin D daily supports the diagnosis of vitamin D-dependent rickets.

Failure of vitamin D alone (despite increased doses) to heal rickets is consistent with vitamin D-resistant rickets. Because this is most commonly an X-linked dominant disorder, evidence of mild rickets/osteomalacia in the child's mother strongly supports the diagnosis of vitamin D-resistant rickets. The treating physician must always evaluate the rachitic child for liver disease and examine the child's medication history. Pancreatitis sufficient to produce hypocalcemia should be blatantly evident and not cause any diagnostic dilemmas. Because of the increasing recognition of vitamin D deficiency in the general population, requests for 25-OHD measurements continue to increase. This increasing prevalence of vitamin D deficiency is likely related to less time spent outdoors, increased coverage of the skin, and increased use of sunscreen to prevent skin cancers (such as melanoma).

Hypercalcemia

Modest hypercalcemia (e.g., total calcium between ~10.5 and ~11.9 mg/dL) may be asymptomatic. Symptomatic hypercalcemia may cause vague complaints such as fatigue, malaise, weakness, depression, apathy, and an inability to concentrate. Other disturbances can include bone pain and gastrointestinal tract disturbances.

The kidney is adversely affected by hypercalcemia through the formation of nephrolithiasis (i.e., kidney stones) with consequent renal colic (i.e., the pain experienced when a stone is passed through the ureter or urethra), mild nephrogenic diabetes insipidus, or renal failure from nephrocalcinosis (i.e., kidney calcification). Other possible findings in hypercalcemia encompass ectopic calcification of the cornea, conjunctiva and blood vessels, cystic bone lesions, pathologic fractures, and osteopenia from bone breakdown. Patients with hyperparathyroidism classically present with "moans" (e.g., CNS complaints), "bones" (e.g., bone pain), "groans" (e.g., gastrointestinal tract disturbances), and "stones" (e.g., kidney stones). Subperiosteal bone resorption noted in the fingers and around the teeth is a classic finding in hyperparathyroidism. There are many potential causes of hypercalcemia (Table 45.3) [52].

TABLE 45.3 Causes of hypercalcemia.

1. Hyperparathyroidism
2. Malignancy
3. Endocrine disorders
4. Granulomatous diseases
5. Drugs
6. Immobilization and other miscellaneous causes

Primary hyperparathyroidism

In the setting of hypercalcemia with a normally functioning parathyroid gland, PTH should be suppressed below the reference interval. On the other hand, hypercalcemia accompanied by a low normal to depressed phosphate concentration and an inappropriately normal to increased PTH is consistent with primary hyperparathyroidism. Elevations in PTH can elevate 1,25-OH_2D concentrations that will also contribute to hypercalcemia. Typically, in hyperparathyroidism, the filtered calcium load is so great that despite increased PTH and its effect on increasing urinary calcium reabsorption, hypercalciuria results.

The differential diagnosis of hyperparathyroidism includes parathyroid adenoma (~80% of cases of primary hyperparathyroidism), parathyroid hyperplasia (15%–20% of cases of primary hyperparathyroidism), and parathyroid carcinoma (<5% of cases of primary hyperparathyroidism). Some sources report that parathyroid carcinoma causes <1% of all cases of primary hyperparathyroidism. In parathyroid adenomas, there is a capsule, possible compression of adjacent tissue and a reduction in the fat content of the adenoma. Adenomas usually occur in a single gland. In parathyroid hyperplasia, all four glands are usually enlarged with a possible reduction in fat cell content of the hyperplastic tissue. The diagnosis is assisted by weighing the four surgically removed parathyroids and finding that each hyperplastic parathyroid gland is increased in weight (normal parathyroid glands weigh <40 mg).

Histologically, the diagnosis of parathyroid carcinoma is not straightforward. Invasion of the capsule supports the diagnosis of carcinoma (versus adenoma), and evidence of metastases is pathognomonic of carcinoma. Besides local invasion, metastatic sites encompass bone, lung, liver, and the adrenal gland. Hypercalcemia in parathyroid carcinoma can be more severe than in other forms of primary hyperparathyroidism with calcium

concentrations possibly exceeding 14−16 mg/dL; PTH also tends to be greatly elevated. In summary, a palpable, firm neck mass in the setting of hypercalcemia is suggestive of a parathyroid carcinoma. If hyperparathyroidism is present and not otherwise explained in a person with a persistently normal calcium concentration, the term "normocalcemic primary hyperparathyroidism" is applied [53,54]. Over time, this condition evolves to display overt hypercalcemia.

Cases of primary hyperparathyroidism are treated surgically. Indications for surgical parathyroidectomy include: (1) serum calcium >12 mg/dL; (2) marked hypercalciuria (e.g., >400 mg/day in adults); (3) any overt clinical manifestation of hyperparathyroidism (e.g., nephrolithiasis); (4) age <50 years; (5) reduced glomerular filtration rate (GFR) not otherwise explained; and (6) cortical bone density radius z-score less than—2 SD below the mean. Indications for surgery in asymptomatic patients with primary hyperparathyroidism include: (1) a markedly increased serum calcium; (2) a history of an episode of life-threatening hypercalcemia; (3) reduced creatinine clearance, presence of kidney stone(s) detected by abdominal radiograph; (4) markedly increased 24-hour urinary calcium excretion; or (5) a substantially reduced bone mass. If any one of these criteria are met (and likely confirmed), parathyroidectomy is indicated in asymptomatic primary hyperparathyroidism [55].

Regarding surgical management of parathyroid adenomas, with surgical removal of the adenomatous parathyroid gland, the three remaining parathyroids can be left in place. Using nuclear medicine sestamibi scans, the adenomatous gland can often be identified preoperatively. If this is the case, in place of a wide-neck incision required for a neck dissection, the surgeon can make a small incision on the side of the neck where the adenoma is anticipated and selectively remove the single adenomatous gland. New technologies for localization of adenomatous parathyroid glands when a sestamibi scan (99mTc-sestamibi parathyroid SPECT) is negative include ^{11}C-methionine positron-emission tomography (PET) with sensitivities and specificities of 86% [56]. ^{18}F-fluorocholine PET/CT exhibits correct rates of 90% or better for the identification of hyperfunctioning parathyroid glands [57].

By measuring PTH intraoperatively and observing a minimum 50% decline in PTH concentrations, successful removal of the parathyroid adenoma is confirmed [58−60]. Such procedures can be performed even in outpatient settings. The baseline PTH should be measured after the induction of anesthesia but before any manipulation of the gland externally or surgically. Squeezing or stimulating a parathyroid gland can result in the rapid release of PTH [61]. Ten minutes after removal of the suspected adenoma, PTH should be reassessed, because PTH has a short half-life of <5 minutes. Intraoperative PTH can be measured in the surgical suite or in a central laboratory if rapid sample transport is coupled with rapid immunochemical analysis. Another proposed approach to identifying a single adenomatous gland (without frozen section diagnosis or intraoperative PTH testing) is to administer sestamibi preoperatively (e.g., within 2.25 hours of surgical removal of tissue) and determine the radioactivity of the surgically removed tissue [59]. This procedure requires validation by other surgical centers.

With parathyroid hyperplasia, usually all of the parathyroids are surgically removed with one-half of one parathyroid gland implanted in the forearm for easy surgical removal in the future should hyperparathyroidism recur [62]. Because the parathyroid glands can exist in ectopic locations, localization by venography with PTH sampling may be necessary following a failed surgical attempt at complete parathyroidectomy. Theoretically, there is no reason why PTH assays during such radiologic procedures cannot be performed in real time using intraoperative technologies.

Lithium can produce hyperparathyroidism in 10%−15% of patients so treated for bipolar disorder. Chronic lithium therapy may raise the set point in the parathyroid gland for negative feedback, thereby eliciting hypercalcemia. This appears to also occur in the kidney tubules as hypocalciuria may be present [63]. Hyperparathyroidism is not uncommon in patients with acromegaly or Zollinger−Ellison syndrome. Hypercalcemia and increased PTH may occur in thiazide-treated patients. Many such patients eventually exhibit primary hyperparathyroidism. However, thiazides must be withdrawn for 2−3 months to determine if hypercalcemia abates, thus attributing increased calcium concentrations to the drug class.

Familial forms of hyperparathyroidism

An asymptomatic, AD familial form of mild, chronic hypercalcemia results from a loss-of-function mutation in the *CaSR* that is expressed on the parathyroid chief cells and on renal tubular cells. Because the parathyroid glands are relatively insensitive to negative feedback from calcium in this case, mild hypersecretion of PTH occurs relative to the patient's calcium. PTH may be increased, although in most cases it is within normal reference limits; phosphate and 1,25-OH_2D concentrations are also normal. Serum calcium usually does not exceed 12 mg/dL. In contrast to other forms of hyperparathyroidism, urinary calcium is not increased, because the renal tubules, which normally express a CaSR, perceive a state of calcium deficiency, leading to high levels of urinary calcium reabsorption and actual hypocalciuria. Therefore this disorder is termed FHH or benign familial hypercalcemia (FHH

type 1; FHH1). Additional causes of the FHH phenotype include inactivating mutations of the G-protein subunit α11 (*GNA11* gene; chr.19p13.3; FHH2) and inactivating mutations of the adaptor-related protein complex 2, sigma 1 subunit (*AP2S1* gene; chr. 19q13.32; FHH3). GNA11 is involved in CaSR signaling as the CaSR employs G-protein-cAMP-adenyl cyclase signaling. AP2S1 plays a role in clathrin-mediated endocytosis of the CaSR, which effects signaling [64]. Mutations in AP2S1 reduce intracellular cycling of CaSR, increasing the expression of the CaSR [65].

Individuals heterozygous for any type of FHH do not display bone pain or gastrointestinal or psychiatric complaints, nor do these individuals develop kidney stones or nephrocalcinosis. However, the clinical differentiation from mild primary hyperparathyroidism may be difficult. In infants homozygous for FHH, severe neonatal hypercalcemia results, which can be life-threatening. Emergency parathyroidectomy may be required to control hypercalcemia in such infants.

Hyperparathyroidism can occur in patients with multiple endocrine neoplasias (MEN). There are two major forms of MEN: MEN1 and MEN2. Furthermore, there are two subtypes of MEN2: MEN2A and MEN2B (note: some authors refer to MEN2A and "MEN2" while referring to MEN2B as "MEN3." This terminology is not preferred because MEN2A and MEN2B result from mutations in the same gene). MEN is inherited as an AD trait, indicating that MEN1 and MEN2 result from germline variants. Mutations in the *MEN1* gene [encoding multiple endocrine neoplasia I gene (MENIN) on chromosome 11q13.1] cause MEN1, and mutations in the *RET* proto-oncogene (on chromosome 10q11.21) cause MEN2. In approximately 5%−10% of persons with the clinical phenotype of MEN1, *MEN1* is not mutated. In some persons that were *MEN1*-negative, heterozygous mutations in the cyclin-dependent kinase inhibitor 1B (*CDNK1B* gene, on chromosome 12p13.1) have been described, resulting in the designation MEN4 [66].

MEN1 is clinically characterized by tumors of the anterior pituitary gland, parathyroid glands, and the pancreatic islets (the "3 Ps"). The parathyroids usually are hyperplastic in MEN1. The most common anterior pituitary tumor is a prolactinoma. Next most common are nonfunctioning pituitary adenomas followed by ACTH-secreting anterior pituitary tumors causing Cushing disease. Less common anterior pituitary tumors secrete growth hormone or TSH.

The most common pancreatic islet cell tumor is a gastrinoma producing a Zollinger−Ellison clinical phenotype. Gastrin serves as a tumor marker for gastrinomas. Next most common are insulinomas (hypoglycemia is discussed in Chapter 33, Evaluation of exocrine pancreatic function). Rare islet cell tumors include somatostatinomas, glucagonomas, and VIPomas. Somatostatinomas, glucagonomas, and VIPomas can produce a secretory, watery, cholera-like diarrhea. Glucagonomas can also produce mild hyperglycemia and a unique skin rash termed "migratory necrolytic epidermolysis."

MEN1 develops when, as part of tumorigenesis, a cell line becomes monosomic for at least a portion of chromosome 11q13.1 containing the *MEN1* gene. The result is that only a defective copy of the tumor-suppressor gene *MEN1* persists. Before loss of heterozygosity, the chromosome 11q13.1 that carried the functional *MEN1* gene prevented cancers. With loss of the normal *MEN1* gene, there is insufficient tumor-suppressor gene activity to prevent cancer. MENIN is a 610−amino-acid, 67-kDa protein. It is expressed throughout the body and is predominantly identified in the nucleus. MENIN has two nuclear localization sequence motifs and interacts with various transcription factors, DNA repair proteins, DNA processing factors, and cytoskeletal proteins including FANCD2, GFAP, junD, NF-kB, NM23beta, nonmuscle myosin heavy chain IIA, PEM, RPA2, and SMAD3. Loss of function of MENIN changes the downstream action of its junD partner from growth suppression to growth promotion. This may explain the neoplastic role of MENIN mutations.

MEN2 manifests as medullary thyroid carcinoma (MTC) and pheochromocytoma in a subset of cases. MTC occurs in ~100% of MEN2 cases, and pheochromocytoma occurs in ~50% of MEN2 cases. MEN2B is distinguished from MEN2A by the presence of neuromas of the tongue, oral mucosa, lips, and eyelids; ganglioneuromatosis of the salivary glands, gall bladder, pancreas, and alimentary canal; evidence of corneal nerve thickening on slit lamp examination; and a marfanoid habitus with skeletal deformities in MEN2B patients. Hyperparathyroidism in MEN2A results from adenomas or hyperplasia. Hyperparathyroidism is possible but uncommon in patients with MEN2B.

MEN2 results from inheritance of a mutated *RET* proto-oncogene. The normal function of the RET proto-oncogene involves signal transduction for glial-derived neurotropic factor (GDNF). The protein that results from the transcription of the *RET* proto-oncogene is a member of the tyrosine-kinase receptor family. There are extracellular domains, a transmembrane domain, and intracellular domains. Normally, GDNF initially binds to a specific cell-surface receptor (GDNF receptor alpha) that then cross-links two copies of the RET proto-oncogene protein (Fig. 45.8). Disulfide bonds form between cysteines present in the paramembrane extracellular domains (encoded by exons 10 and 11) of the two RET proto-oncogene receptors. With cross-linking of the two RET proto-oncogene proteins, tyrosine kinase activity develops within the two intracellular domains, allowing for downstream autophosphorylation and phosphorylation.

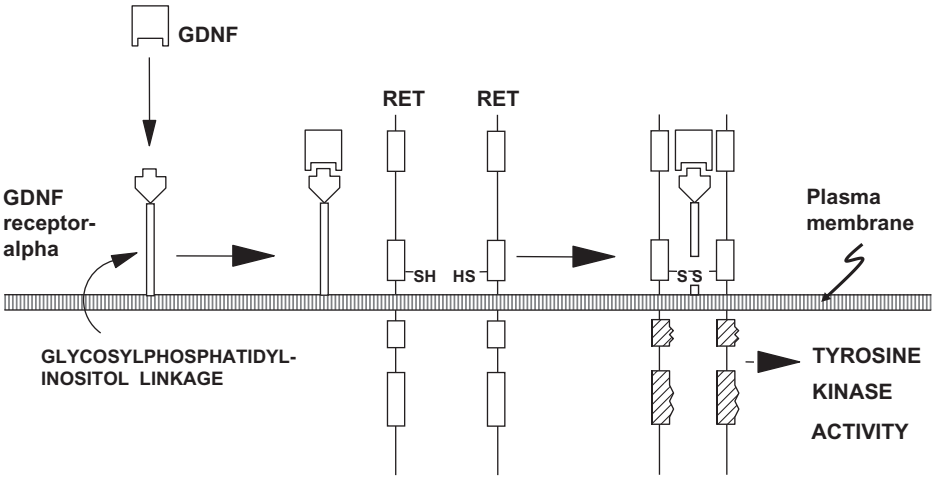


FIGURE 45.8 The normal function of the *RET* proto-oncogene is to interact with glial cell-derived neurotrophic factor receptor-alpha after it has bound GDNF. Two *RET* proto-oncogene proteins cross-link upon binding the GDNF-GDNF receptor alpha complex to initiate signal transduction via the development of tyrosine kinase activity in *RET*.

The locations of mutations in the RET proto-oncogene differ between the two forms of MEN2. In MEN2A, the mutations are located in the extracellular, paramembrane, cysteine-rich domains that foster spontaneous dimerization in the absence of ligand binding. In MEN2B, there is a gain-of-function mutation in either of the two cytoplasmic tyrosine kinase domains (Fig. 45.9). Activation then occurs in the absence of RET proto-oncogene protein dimerization. Tyrosine kinase domain 1 is encoded by exons 13, 14, and 15, whereas tyrosine kinase domain 2 is encoded by exon 16. MEN2B is caused by mutations in exon 16 more commonly than in exons 13, 14, or 15. Of interest, there is one report of the coexistence of MEN2 and APS type 2 [67]. However, the patient described did not have adrenal disease, suggesting that the patient's APS was not APS type 2.

If there is a family history of MEN2 and the proband (a person serving as the starting point for the genetic study of a family) has been studied on a molecular basis, relatives can be specifically studied for the same mutation. However, if the proband has not been studied on a molecular basis, then relatives will require study of their entire RET proto-oncogene for mutations in the exons known to harbor mutations causative of MEN2. Once one mutation is recognized, other first-degree relatives can be studied selectively for this mutation. Specific MEN2 mutations can help guide therapy (e.g., in the timing of elective thyroidectomy in MEN2) [68,69].

Regarding MEN4, *CDNK1B* encodes the 196 amino-acid CDK1 p27Kip1 that is activated by H3K4 methylation [70]. Failure of normal CDK1 p27Kip1f function appears to allow activation of cyclin E-CDK2 or cyclin D-CDK4 complexes, permitting cell cycle progression at the G1 cell-cycle phase (Fig. 45.10). Progression to S phase involves the *RB1* (retinoblastoma 1) gene [encoding the RB Transcriptional Corepressor 1 (chr. 13q14.2)] modifying the gene from a hypophosphorylated state to a hyperphosphorylated state. A novel role for CDKN1B as a tumor susceptibility/suppressor gene is supported by the detection of CDKN1B mutations (either germ line or somatic) in some patients with breast cancer, nonfamilial primary hyperparathyroidism, lymphoma, and small intestinal neuroendocrine tumors.

Still another familial cancer syndrome involving the parathyroid glands is the hyperparathyroidism-jaw tumor syndrome (HPT-JT) [71]. Hyperparathyroidism most often results from a parathyroid adenoma but can result from a parathyroid carcinoma. Other tumors can occur (e.g., fibroosseous tumors of the jaw or maxilla) but uterine tumors are particularly common in women with HPT-JT. This AD condition results from HRPT2 [hyperparathyroidism 2 homolog (human); chromosome 1q24-q32] loss-of-function mutations. HRPT2 has also been named CDC73 (cell division cycle 73; a.k.a. parafibromin). CDC73 normally functions as a tumor suppressor gene. HPT-JT may occur sporadically where such CDC73 mutations are believed to occur somatically. HPT-JT should be considered in cases of early-onset hyperparathyroidism to inform the choice of surgical management (e.g., removal of all parathyroids in HPT-JT) [72]. Familial isolated hyperparathyroidism may represent incomplete expression of FHH, MEN1, and HPT-JT, while some patients have an activating glial cells missing homolog 2 (*GCM2*; chr. 6p24.2) mutation. GCM2 is involved in parathyroid gland development. In summary, there are now at least eight familial syndromes potentially causing hyperparathyroidism (Table 45.4) [73].

Another rare form of primary hyperparathyroidism is autoimmune hypercalcemia. This condition results from the production of a CaSR autoantibody that blocks the

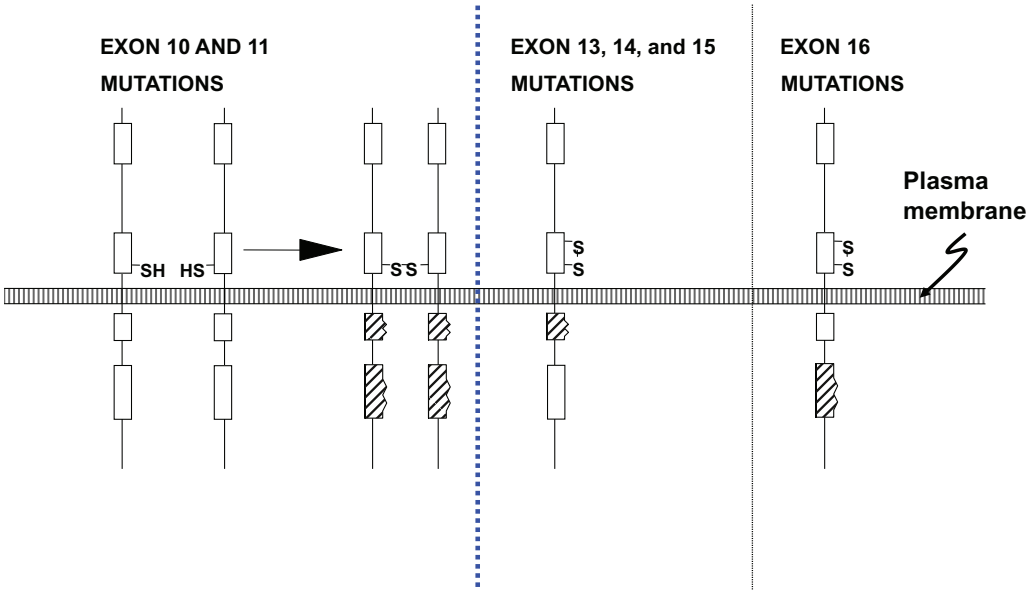


FIGURE 45.9 MEN2A is caused by mutations in exons 10 or 11 that allow spontaneous homodimerization of RET in the absence of the GDNF-GDNF receptor alpha complex. In MEN2B, mutations in exons 13, 14, or 15 produce gain-of-function mutations in the either of the tyrosine kinase domains.

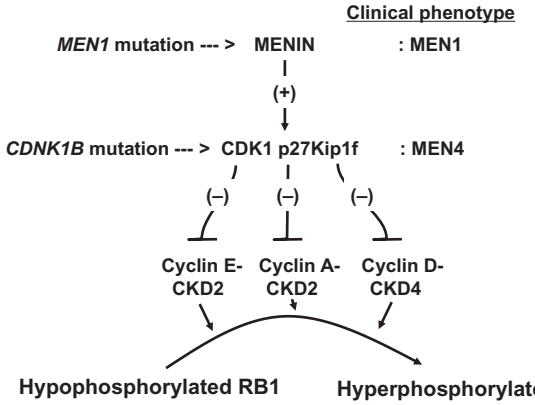


FIGURE 45.10 MENIN normally stimulates CDK1 p27Kip1f. In turn, CDK1 p27Kip1f suppresses cyclin/chronic kidney disease (CKD) complexes. Cyclin/CKD complexes advance cells through the G1 stage by phosphorylating RB1 (Retinoblastoma transcriptional corepressor 1; chr.13q14.2). Mutations in *MEN1* cause MEN1. Mutations in *CDNK1B* that encode CDK1 p27Kip1f cause MEN4.

receptor's perception of ionized calcium, producing increased and pathologic PTH secretion. Similar to FHH and lithium-induced hyperparathyroidism, hypocalciuria can be present.

Secondary hyperparathyroidism

Although not usually a cause of hypercalcemia, secondary hyperparathyroidism is included because of its adverse effect on bone biology. Secondary hyperparathyroidism is a normal response to hypocalcemia. Deficient vitamin D action was also described previously as a cause of secondary hyperparathyroidism.

Secondary hyperparathyroidism is a major consequence of chronic renal failure. Several factors contribute to secondary hyperparathyroidism in chronic renal failure. First, the kidneys' failure to excrete phosphate produces hyperphosphatemia. Hyperphosphatemia then lowers calcium concentrations with the development of ectopic calcification. Hyperphosphatemia also inhibits renal 1 alpha-hydroxylase activity, impairing the conversion of 25-OHD to 1,25-OH_2D in the renal tubule. Declining 1,25-OH_2D concentrations impair calcium absorption from the diet. In addition, renal tubular disease or destruction can also lead to a deficiency of 1,25-OH_2D. With a decline in calcium concentrations, PTH is secreted (i.e., secondary hyperparathyroidism). The calcium consequently rises back into the lower reference interval but at the expense of bone destruction from bone resorption. Such renal bone disease is histologically referred to as "osteitis fibrosa cystica." Secondary hyperparathyroidism of renal failure is treated by administering vitamin D (usually the active form of vitamin D: 1,25-OH_2D) and oral phosphate binders that inhibit phosphate absorption from the gastrointestinal tract. Routine monitoring is undertaken through measurements of calcium, phosphate, alkaline phosphatase, and PTH.

As noted above, in general, true-intact PTH assays are not superior to intact PTH assays [74,75]. In the cases of patients with secondary hyperparathyroidism undergoing parathyroidectomy for a pathologic excess of PTH, there are data suggesting that PTH measured intraoperatively

TABLE 45.4 Familial causes of hyperparathyroidism.

Disease	Gene mutated
FHH1	*CASR*
FHH2	*GNA11*
FHH3	*AP2S1*
MEN1	*MEN1*
MEN2A	*Ret proto-oncogene*
MEN2B	*Ret proto-oncogene*
MEN4	*CDKN1B*
HPT-JT	*HRPT2 (CDC73)*
Familial isolated hyperparathyroidism	*GCM2*; see text also

AP2S1, Adaptor-related protein complex 2, sigma 1; *CaSR*, calcium-sensing receptor; *CDKN1B*, cyclin-dependent kinase inhibitor 1B; *FHH*, familial hypocalciuric hypercalcemia; *GCM2*, glial cells missing homolog 2; *GNA11*, G-protein subunit α11; *HPT-JT*, hyperparathyroidism-jaw tumor; *MEN*, multiple endocrine neoplasia.

with the whole (1−84) PTH assays will decline more rapidly as than the PTH that is measured with the traditional intact (7−84) PTH assays [76]. The degree to which this proves valuable in clinical practice is unproven.

Tertiary hyperparathyroidism

In patients with secondary hyperparathyroidism of renal failure, renal transplantation with subsequent normal production of 1,25-OH_2D and normal phosphate excretion should permit the return of normal parathyroid function. However, if hyperparathyroidism continues despite renal transplantation, the parathyroid glands have become hyperplastic to the point that they are autonomous. This is termed "tertiary hyperparathyroidism." Hypercalcemia can occur. Like primary hyperparathyroidism, surgical removal of the glands is required. Typically, 3.5 parathyroids are surgically removed and the remaining one-half of 1 parathyroid is implanted in the forearm. Forearm placement of this parathyroid tissue will allow easy surgical removal should hyperparathyroidism recur in this autotransplanted tissue.

Malignancy

Excluding hyperparathyroidism, malignancy is the most common cause of severe and/or persistent hypercalcemia. There are several mechanisms, whereby cancer can cause increased calcium concentrations. One mechanism involves the secretion of PTHrP. PTHrP has actions similar to PTH. However, PTHrP is not detected in the PTH immunoassays. A separate assay specific for PTHrP is available. Three isoforms of PTHrP (139, 141, and 173 amino acids) with N-terminal homology to PTH are encoded by the PTHrP gene (PTH like hormone; *PTHLH*; chr. 12p11.22). Another term for this form of paraneoplastic hypercalcemia is "humoral hypercalcemia of malignancy."

The second mechanism of hypercalcemia in malignancy is local bone destruction by the secretion of cytokines or other osteoclast-activating factors by tumor metastases. This is referred to as "local osteolytic hypercalcemia of malignancy." Numerous tumors cause this type of hypercalcemia including multiple myeloma, leukemias, lymphomas, and breast, colon, and lung cancers. Wilms tumor in children is another cause of this form of hypercalcemia. One interesting feature of multiple myeloma is that alkaline phosphatase measurements are not elevated (assuming a fracture is not present) [77]. Osteoblasts also appear to be suppressed in myeloma-bone lesions by cytokines (e.g., IL-3, IL-7, TNF-α, dickkopf 1, and soluble-frizzled receptor-like proteins 2 and 3) and via surface contact from stromal cells (e.g., activin A and growth factor-independent 1 transcriptional repressor).

A third form of cancer-related hypercalcemia has been observed in certain cases of lymphoma. The lymphoma appears to be able to convert 25-OHD into 1,25-OH_2D. Increased 1,25-OH_2D then produces hypercalcemia through increased calcium absorption from the gut and bone. This is similar to hypercalcemia resulting from granulomas. Rarely, PTH is produced ectopically.

Endocrine disorders

Both hypothyroidism and hyperthyroidism have been associated with hypercalcemia. In hypothyroidism, the proposed mechanism is prolonged action of vitamin D because of slowed metabolism. Increased bone turnover with bone loss is the proposed cause of hypercalcemia in hyperthyroidism. Alkaline phosphatase may also be increased in such cases of thyrotoxicosis.

Although glucocorticoids acutely suppress bone turnover and can lower calcium, glucocorticoid deficiency (e.g., Addison disease) can lead to increased calcium in the setting of Addisonian crisis. Volume contraction and hemoconcentration may also contribute to the hypercalcemia by elevating the relative concentration of albumin, the major calcium-binding protein. VIPomas can produce hypercalcemia independently of hyperparathyroidism; however, the mechanism causing hypercalcemia is unclear [78].

Granulomatous diseases

Any form of granulomatous disease can cause hypercalcemia through the production of 1,25-OH_2D. Normally, the major site of conversion of 25-OHD into 1,25-OH_2D is the renal tubule. However, in granulomatous diseases, the macrophages may acquire the ability to convert 25-OHD into 1,25-OH_2D. The list of possible granulomatous

diseases includes sarcoidosis, histoplasmosis, coccidioidomycosis, cryptococcosis, leprosy, berylliosis, and silicosis.

Drugs

Vitamin D excess causes pathologic elevations in both calcium and phosphate through increased intestinal absorption and increased bone mobilization [79]. Because vitamin D is fat-soluble, toxicity can be prolonged for weeks to months. Vitamin A excess has been reported to cause hypercalcemia. In some of these cases, there is coexistent vitamin D excess. Vitamin A itself in high doses appears to be able to cause bone resorption. Thiazides increase calcium absorption from the urine ultrafiltrate but rarely cause significant hypercalcemia. Because of thiazides' effects on increasing urinary calcium reabsorption, thiazides were studied as therapeutic agents intended to benefit osteoporosis. However, they were clinically ineffective.

Toxic doses of PTH administered in the treatment of osteoporosis can cause hypercalcemia. An FDA-approved drug for the treatment of osteoporosis is teriparatide, which contains the first 34 amino acids of PTH. Studies in rats demonstrated an increased risk of developing osteosarcoma.

Immobilization and other miscellaneous causes of hypercalcemia

Especially in growing teenagers, prolonged immobilization will raise calcium as calcium is released from the skeleton. Although weightlessness in outer space does not cause hypercalcemia, accelerated bone resorption in weightless conditions will be a major hazard of extended-duration space missions or travel to distant planets such as Mars.

Milk-alkali syndrome resulting from large intakes of calcium and alkali leads to both increased calcium absorption (and subsequent hypercalcemia) and metabolic alkalosis. The typical patient who develops milk-alkali syndrome is treating him or herself with milk and HCO_3^- for peptic ulcer disease. In hospitalized patients, total parenteral nutrition must be considered as a cause of hypercalcemia.

Approach to hypercalcemia

Once dehydration or hemoconcentration has been excluded as causes of hypercalcemia through measurements of albumin or ionized calcium (the preferred measurement), PTH is measured to differentiate hyperparathyroidism from all other causes of hypercalcemia (Fig. 45.11). Phosphate and alkaline phosphatase should be measured as well. If hyperparathyroidism is documented (e.g., the PTH is inappropriately normal or elevated in the setting of true hypercalcemia), familial causes should be excluded. The evaluation should include an assessment of the 24-hour urinary calcium excretion.

When PTH is suppressed in the face of hypercalcemia, non−PTH-dependent hypercalcemia is pursued. A thorough search for malignancy must be undertaken. Measurement of serum phosphate concentrations is informative in cases of non−PTH-dependent hypercalcemia. For example, with normal renal function, the presence of hypercalcemia and hyperphosphatemia is suggestive of vitamin D toxicity or granulomatous disease. Assuming

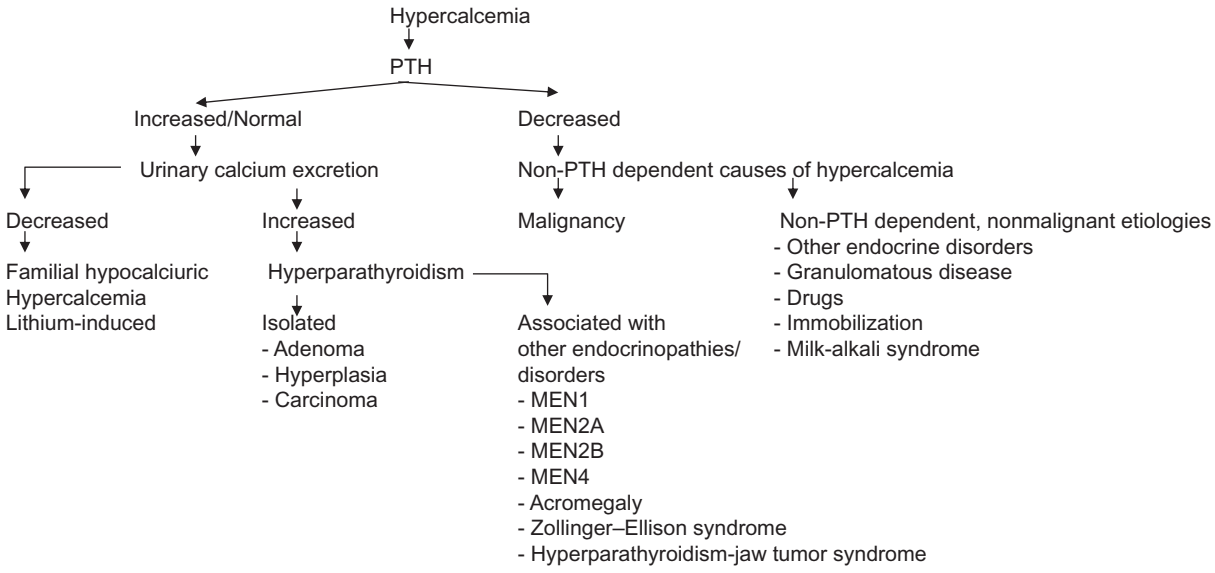


FIGURE 45.11 Diagrammed is an approach to the evaluation of hypecalcemia. See the text for details.

that malignancy is ruled out, endocrine, drug, and various miscellaneous causes of hypercalcemia are then stepwisely investigated.

Mutations in the parathyroid hormone/parathyroid hormone-related peptide receptor

PTH/PTHrP receptor mutations cause rare genetic bone disorders. A gain-of-function mutation in the PTH/PTHrP receptor results in Jansen metaphyseal chondrodysplasia [80]. This AD disorder is characterized by short-limbed disproportionate short stature and abnormal regulation of enchondral bone formation, hypercalcemia independent of either PTH or PTHrP, and increased bone turnover. PTH/PTHrP receptor mutations have also been reported in a few cases of enchondromatosis (Ollier disease) in the absence of hypercalcemia. Isocitrate dehydrogenase [NADP(+)] 1, cytosolic (IDH1; chr. 2q34), and isocitrate dehydrogenase [NADP(+)] 2, mitochondrial (IDH2; chr. 15q26.1) have been reported in some patients with Ollier disease [81].

Loss-of-function mutations in the PTH/PTHrP receptor are reported to cause Blomstrand chondrodysplasia. This rare, AR disorder is manifested in early death and advanced bone maturation with accelerated chondrocyte differentiation.

Calcitonin

Calcitonin is a product of the C-cells (a.k.a. parafollicular cells) of the thyroid gland. Less than 0.1% of the thyroid epithelial mass is C cells. The calcitonin gene family has five members: *CALC-I* encodes calcitonin and calcitonin gene-related peptide-I, *CALC-II* encodes calcitonin gene-related peptide-II, *CALC-III* is a pseudogene; *CALC-IV* encodes amylin (a.k.a. islet-associated polypeptide), and *CALC-V* encodes adrenomedullin. Preprocalcitonin is cleaved to a signal peptide plus procalcitonin (116 amino acids; 12.8 kDa). Procalcitonin is further cleaved to aminoprocalcitonin (N-terminal procalcitonin) plus conjoined calcitonin: calcitonin carboxypeptide-I. Cleavage of calcitonin: calcitonin carboxypeptide-I yields immature calcitonin plus calcitonin carboxypeptide-I. Finally, immature calcitonin (33 amino acids; terminal glycine) loses one amino acid and is modified to mature amidated calcitonin (32 amino acids; terminal amidated proline).

Calcitonin lowers serum calcium by stimulating bone accretion. Calcitonin plays only a minor physiologic role in humans: complete thyroidectomy has no adverse effect on BMD or bone strength if the subject is appropriately replaced with thyroid hormone. In clinical medicine, calcitonin is measured as a tumor marker for MTC. Procalcitonin is a marker used for the diagnosis of bacterial infections [82]; however, procalcitonin measurements have no role in endocrine testing.

Because MTC is derived from the C cells of the thyroid gland, elevated calcitonin concentrations serve as a tumor marker for MTC. Most experts consider basal calcitonin concentrations ≤ 20 pg/mL as normal. Such values do not require further study. A basal calcitonin ≥ 100 pg/mL is frankly elevated. If the basal calcitonin is >20 pg/mL but less than 100 pg/mL, calcitonin should be measured following stimulation, e.g., following the IV injection of calcium. In this test, 2 mg/kg of elemental calcium is infused intravenously as calcium gluconate over 50–60 seconds. Calcitonin is then measured at +1, +3, +5, and +10 minutes. Previously, pentagastrin (dose: 0.5 mcg/kg IV push over 5–10 seconds) was used as a calcitonin-secretagogue alone or together with IV calcium. However, pentagastrin is not presently available for diagnostic testing in the United States. An elevated calcitonin (e.g., ≥ 100 pg/mL) in a patient with a thyroid mass (e.g., nodule or nodules) is presumptive evidence of MTC. However, the final diagnosis of MTC is dependent on the gross and microscopic examination of the surgically excised tissue.

As noted above, MTC can occur sporadically (75% of cases) or as part of a familial cancer syndrome (25% of cases). Because the likelihood of MTC occurring in MEN2 is nearly 100% in affected subjects, prophylactic complete thyroidectomy should be performed. The specific *RET* proto-onocgene mutation can inform the medical staff as to how aggressive they should be in thyroidectomy (e.g., early versus later thyroidectomy). The preoperative diagnosis of familial cancer syndromes associated with MTC can be performed by testing for mutations in the *RET* proto-oncogene. The preoperative diagnosis specifically of MTC can be made by fine needle aspiration biopsy of a thyroid lesion or, more effectively, by measuring calcitonin as described above. Once thyroidectomy has been accomplished, calcitonin can be measured serially as a tumor marker.

Calcitonin measurements are available as radioimmunassays and as two-site immunoenzymatic or immunoradiometric assays. The differential diagnosis of elevated calcitonin includes MTC, C-cell hyperplasia (a precursor of MTC), certain nonthyroidal cancers (e.g., small-cell carcinomas, oat cell carcinomas, and other malignancies), acute or chronic renal failure, hypercalcemia, hypergastrinemia and other gastrointestinal disorders, and pulmonary disease.

Phosphate

Traditionally, phosphate has been viewed as subordinate to calcium concentrations: the body primarily regulates calcium and in the process of regulating calcium, phosphate is regulated. However, with the discovery that FGF23 has a role in phosphate homeostasis, independent control of phosphate has emerged as an important issue [83].

FGF23 is derived from bone, and together with the membrane bound protein Klotho, FGF23 has the following actions: (1) induction of renal phosphaturia by reducing the activity of Npt-2 (NaPi-2a and NaPi-2c; note: NaPi-2b is regulated by calcitonin); (2) reduced expression of renal 1 alpha-hydroxylase; and (3) increased clearance of 1,25-OH_2D via stimulation of 24-hydroxylase. Thus the overall effect of FGF23 is to lower plasma phosphate concentrations. Klotho associates with the receptor for FGF23. Klotho is encoded by the *KL* gene on chromosome 13q13.1. Klotho is composed of 1012 amino acids and weighs ~116 kDa.

Plasma phosphate concentrations represent the balance of phosphate entering the body through the gastrointestinal tract or IV route, the distribution of phosphate inside and outside of cells and its deposition into bone, and the renal clearance of phosphate (Fig. 45.12). Total body phosphate is predominantly controlled by controlling the degree of phosphaturia, with both PTH and FGF23 increasing phosphate excretion.

Elevations in phosphate occur for the following reasons: (1) increased phosphate burden [e.g., increased phosphate intake from any source (oral, rectal, cutaneous, parenteral) or increased phosphate absorption from the gastrointestinal tract]; (2) phosphate redistribution within the body (e.g., increased cellular release of phosphate from cell injury, dysfunction or cell lysis, or increased bone turnover or release from bone); and (3) decreased renal phosphate clearance (e.g., reduced renal excretion or increased renal tubular reabsorption) (Table 45.5). The causes of hypophosphatemia are essentially the inverse of hyperphosphatemia: (1) inadequate intake; (2) phosphate redistribution into cells; and (3) excessive renal loss (Table 45.6).

Magnesium

Although magnesium is the fourth most abundant cation in the body (following sodium, potassium, and calcium) and is the second most abundant intracellular cation (following potassium), there is no "primary" hormone that controls circulating magnesium concentrations. Nevertheless, magnesium is important in cellular metabolism. Acutely, PTH appears to have some regulatory effect on magnesium as acute hypoparathyroidism will reduce magnesium concentrations. On the other hand, chronic hypomagnesemia impairs PTH release and reduces target tissues' responses to PTH.

The most common cause of hypermagnesemia in clinical practice is iatrogenic: the intended infusion of magnesium sulfate in the treatment of preeclampsia or eclampsia in pregnant women. Infants born to magnesium-treated mothers can also be hypermagnesemic. Other causes of hypermagnesemia include excessive oral intake, increased gastrointestinal tract absorption (e.g., hypomotility and milk-alkali syndrome), cellular release (e.g., tumor lysis syndrome and rhabdomyolysis), and contraction of the circulating blood volume.

Transient hypomagnesemia is not uncommon in severe acute or critical illnesses. This may result from transient hypoparathyroidism that is self-limited but also common in such states. The most common causes of chronic hypomagnesemia are malabsorption, malnutrition, and renal magnesium wasting from drugs (e.g., loop or thiazide diuretics; cisplatin; cyclosporine, or tacrolimus), endocrine disorders (e.g., primary hyperaldosteronism, hypoparathyroidism, and hyperthyroidism) or osmotic diuresis.

Markers of bone turnover

Osteoporosis is a major medical problem for postmenopausal women and the elderly in general [84]. The costs associated with the consequences of osteoporosis run into the billions of dollars per year. There is controversy concerning the role of laboratory testing in the assessment of bone turnover during osteoporosis treatment [50]. If any

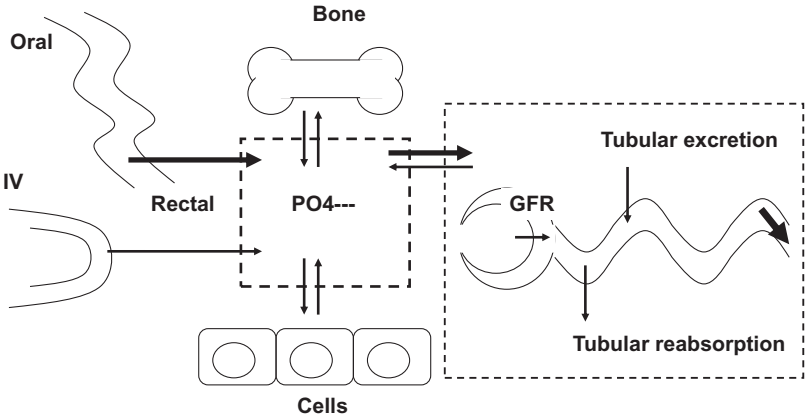


FIGURE 45.12 Phosphate is plentiful in normal diets and phosphate balance is maintained by the urinary excretion of phosphate. Appropriate excretion is a function of a normal glomerular filtration rate and tubular excretion, and the lack of excessive tubular reabsorption of phosphate. Phosphate exchanges with pools in the cells and bone.

TABLE 45.5 Causes of hyperphosphatemia.

Increased phosphate burden
Increased phosphate intake
Oral/rectal
- Noncompliance to phosphate binders in renal insufficiency
- Phosphate salts (e.g., oral/rectal laxatives and enemas)
Cutaneous: white phosphorous burns
Parenteral
Transfusion of outdated blood
Liposomal amphotericin B
Intravenous phosphate
Hyperalimentation (including lipid administration)
Increased phosphate absorption from the gastrointestinal tract
Vitamin D excess[a] (multiple causes including sarcoidosis)
Milk-alkali syndrome
Phosphate redistribution within the body
Increased bone turnover or release from bone
Thyrotoxicosis
Prolonged immobilization
Glucocorticoid withdrawal or deficiency
Increased cellular release of phosphate: cell injury, dysfunction, or cell lysis
Cancer (tumor lysis syndrome)
Trauma (including crush injures; rhabdomyolysis)
Burns
Ischemia or infarction (e.g., involving the bowel)
Thermal injury (heat or hypothermia)
Infection
Shock
Diabetic ketoacidosis, lactic acidosis, acute respiratory acidosis, and alcoholic ketoacidosis
Hemolysis
Inborn errors (e.g., malignant hyperthermia)
Exhaustive exercise
Decreased renal phosphate clearance
Increased tubular reabsorption of phosphate
Growth hormone excess (e.g., acromegaly)
Bisphosphonates
Tumoral calcinosis (mutations in FGF23 or uridine diphosphate-*N*-acetyl-alpha-D-galactosamine:polypeptide *N*-acetylgalactosaminyltransferase 3 or Klotho)
Decreased renal excretion:
Reduced glomerular filtration rate
Renal disease (acute or chronic)
Normal glomerular filtration rate: reduced tubular excretion of phosphate
Hypoparathyroidism
Pseudohypoparathyroidism
Alkalosis (acquired PTH resistance)

FGF23, Fibroblast growth factor-23; *PTH,* parathyroid hormone.

[a]*Also causes increased bone turnover.*

bone marker is going to be used to monitor the rate of bone turnover, it should be measured at baseline prior to the institution of any antiresorptive therapy. This allows each patient to serve as their own control.

Bone markers are breakdown products of collagen or are products of bone cells such as osteoblasts (e.g., BAP; a.k.a. skeletal alkaline phosphatase and osteocalcin) [85].

Osteocalcin (encoded by the bone gamma-carboxyglutamate protein gene on chr. 1q22) is a cytosolic 5.7-kDa calcium binding protein (49 amino acids). Osteocalcin is the major noncollagen protein found in bone and represents ~1% of total protein. Via the action of vitamin-K–dependent carboxylases, amino acids are modified into gamma-carboxy glutamyl residues that can

TABLE 45.6 Causes of hypophosphatemia.

Inadequate intake
Inadequate dietary intake
Malabsorption of intestinal phosphate
Phosphate binding by the antacids
Vitamin D deficiency (inadequate absorption)
Phosphate redistribution into cells
Acute respiratory alkalosis or hyperventilation
Insulin administration
Epinephrine injection
Cytokine-induced
Hungry bone syndrome (healing bone)
Vitamin D replenishment
Excessive loss
Renal phosphate wasting (genetic or acquired renal disorders)
Rickets with secondary hyperparathyroidism
Familial forms of hypophosphatemic rickets
Diseases of or injury to proximal renal tubule: heavy metal poisoning, various drugs, and monoclonal gammopathies
Drug-induced renal phosphate wasting (e.g., loop diuretics)
Increased glomerular filtration rate (including diuretics and osmotic agents)
Alkalosis
Tumor-induced osteomalacia (FGF23 secretion)
Following kidney transplantation (transient)

bind calcium or hydroxyapatite. Osteocalcin has a short half-life (~5 minutes). Between 10% and 30% of osteocalcin produced by osteoblasts is released extracellularly.

Increased osteocalcin or BAP levels signal an elevated rate of bone formation. Because bone building and bone breakdown are interrelated and because in the postmenopausal woman and in the elderly bone breakdown usually exceeds bone formation, elevated BAP or osteocalcin are undesirable suggesting an overall decline in BMD. In addition to osteoporosis, other causes of elevated BAP or osteocalcin include osteomalacia, rickets, renal osteodystrophy, Paget disease of bone, hyperparathyroidism, acromegaly, and thyrotoxicosis. Increased BAP also occurs with osteolytic metastases.

With bone resorption, osteoclasts degrade type 1 collagen. Type 1 collagen, which represents ~90% of the organic matrix of bone, is composed of three individual procollagen chains: two α1 procollagen chains and one α2 procollagen chain. The procollagen chains are composed of three regions: the amino-terminal propeptide, the helical domain and the carboxy-terminal propeptide. For the α1 procollagen chains, the amino-terminal propeptides include a globular domain. The procollagen chains are modified following translation with the addition of sugars and formation of intrachain disulfide bonds and the addition of hydroxyl groups on proline and lysine residues.

A triple helix of the three procollagen chains (two α1 and one α2) is formed within osteoblasts (Fig. 45.13) and is then secreted outside the cell (Fig. 45.14). The secreted

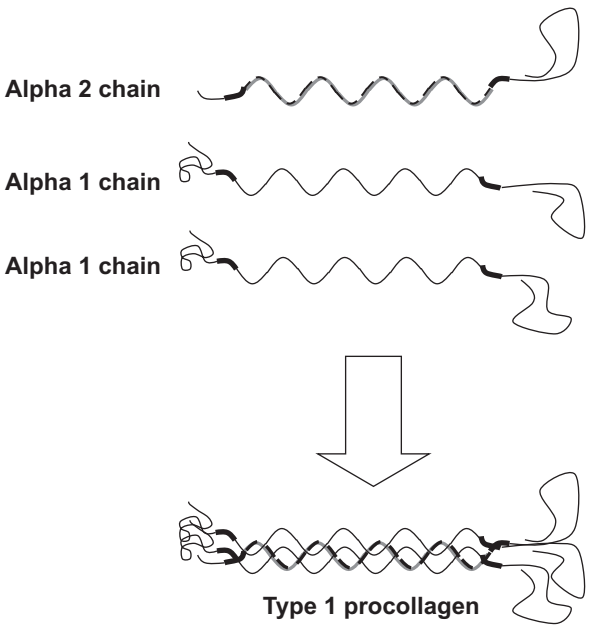


FIGURE 45.13 Within osteoblasts, three procollagen chains (two α1 and one α2) from a triple helix constituting type 1 procollagen.

type 1 procollagen undergoes removal of propeptides via the action of procollagen amino proteinase and carboxy proteinase. This releases the amino and carboxyl terminal propeptides yielding the mature triple helix type 1 collagen (Fig. 45.14). Cross-linking of the collagen triple helices occurs extracellularly (Fig. 45.15). The pyridinium

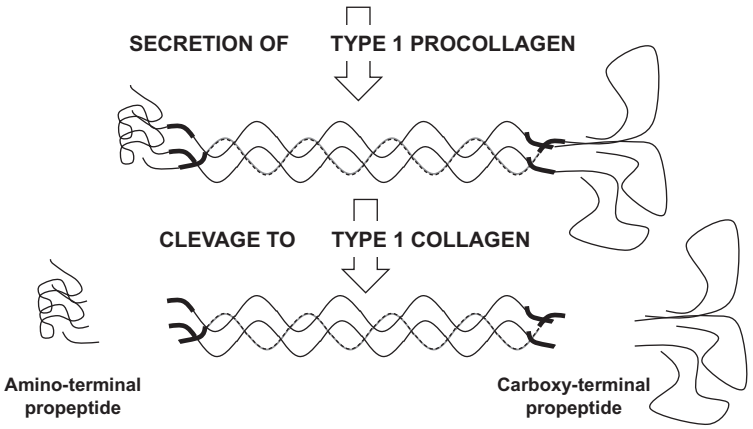


FIGURE 45.14 The secreted triple helix of type 1 procollagen undergoes removal of propeptides via the action of procollagen amino proteinase and carboxy proteinase releasing the aminoterminal and carboxyterminal propeptides to form type 1 collagen.

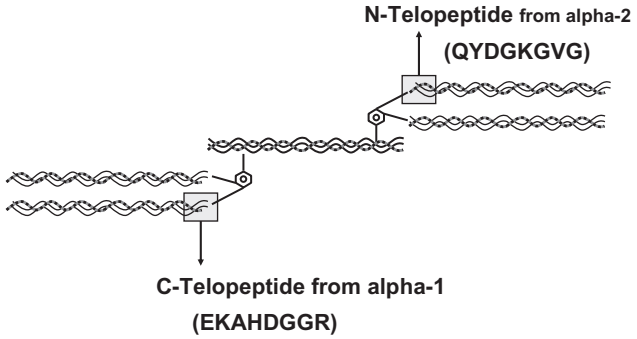


FIGURE 45.15 The cross-linking of the type 1 collagen triple helices occurs extracellularly. The pyridinium cross-links (i.e., the hexagons in the figure) bridge 2 N-telopeptide or 2 C-telopeptide portions of two strands to the helix of another collagen strand. When type 1 collagen is broken down as part of bone resorption, N-terminal telopeptides (N-telopeptides) and C-terminal telopeptides (C-telopeptides) are released.

cross-links (i.e., pyridinoline or deoxypyridinoline) bridge 2 N-telopeptide or 2 C-telopeptide portions of two strands of 2 helices. Lysyl oxidase converts 3-amino acids (i.e., lysines and hydroxylysines) into one trivalent amino acid (deoxypyridinoline and pyridinoline) producing the cross-linking. These cross-linked triple helices account for the very high tensile strength of collagen.

As part of bone resorption when collagen is degraded, pyridinolines, deoxypyridinolines, C-terminal telopeptides, and N-terminal telopeptides are released (Fig. 45.15). An immunoassay for serum C-telopeptides recognizes the following amino acid sequence within the $\alpha 1$ molecule C-telopeptide: EKAHDGGR. An immunoassay for N-telopeptides recognizes the following amino acid sequence within the $\alpha 2$ chain N-telopeptide: QYDGKGVG.

N-telopeptides can be measured in serum or urine. C-telopeptides are measured in serum. Using high-pressure liquid chromatography, ELISA, or radioimmunoassay techniques, deoxypyridinolines or pyridinolines are usually measured in timed urine collections. Deoxypyridinoline is believed to be a more specific marker of bone collagen turnover than pyridinoline.

Whereas monitoring BMD and bone markers in the treatment of osteoporosis is logical, there are few data to demonstrate improved compliance or outcome with treatment when either BMD or bone markers are serially measured [67]. If BMD decreases or bone marker concentrations increase, it is not recommended that therapy for osteoporosis be discontinued. Thus our discussion on the clinical use of bone markers ends as it began: the effectiveness of bone marker monitoring is controversial [86].

Summary

Bone is a tissue scaffolding that grows as we progress from infants to children to teenagers to adults. Bone is amazing in its complexing and ability to self-heal (within certain limits). However, bone is just one aspect of calcium and phosphate biology that demands our attention. Understanding PTH, vitamin D, FGF23, and the many other factors that help maintain homeostasis forms the basis for medically related practices. Most medical activities relating to calcium and PTH biology and disease diagnosis and management require appropriate and prudent laboratory testing.

References

[1] J.J. Brown, M.R. Hynes, J.H. Wible Jr., Measurement of serum calcium concentration after administration of four gadolinium-based contrast agents to human volunteers, Am. J. Roentgenol. 189 (6) (2007).

[2] L. D'Souza-Li, The calcium-sensing receptor and related diseases, Arq. Bras. Endocrinol. Metab. 50 (4) (2006) 628–639.

[3] P. D'Amour, J.H. Brossard, L. Rousseau, L. Roy, P. Gao, T. Cantor, Amino-terminal form of parathyroid hormone (PTH) with immunologic similarities to hPTH(1-84) is overproduced in primary

and secondary hyperparathyroidism, Clin. Chem. 49 (2003) 2037–2044.

[4] P. Gao, S. Scheibel, P. D'Amour, M.R. John, S.D. Rao, H. Schmidt-Gayk, et al., Development of a novel immunoradiometric assay exclusively for biologically active whole parathyroid hormone 1-84: implications for improvement of accurate assessment of parathyroid function, J. Bone Min. Res. 16 (2001) 605–614.

[5] M. Inaba, K. Nakatsuka, Y. Imanishi, M. Watanabe, Y. Mamiya, E. Ishimura, et al., Technical and clinical characterization of the Bio-PTH (1-84) immunochemiluminometric assay and comparison with a second-generation assay for parathyroid hormone, Clin. Chem. 50 (2004) 385–390.

[6] H. Reichel, A. Esser, H.J. Roth, H. Schmidt-Gayk, Influence of PTH assay methodology on differential diagnosis of renal bone disease, Nephrol. Dial. Transpl. 18 (2003) 759–768.

[7] I.B. Salusky, H. Juppner, New PTH assays and renal osteodystrophy, Pediatr. Nephrol. 19 (2004) 709–713.

[8] I.B. Salusky, W.G. Goodman, B.D. Kuizon, J.R. Lavigne, R.J. Zahranik, B. Gales, et al., Similar predictive value of bone turnover using first and second-generation immunometric PTH assays in pediatric patients treated with peritoneal dialysis, Kidney Int. 63 (2003) 1801–1808.

[9] R. Savoca, A. Bock, M.E. Kraenzlin, H.R. Schmid, A.R. Huber, An automated 'bio-intact' PTH assay: a step towards standardization and improved correlation with parathyroid function in renal disease, Clin. Chim. Acta 343 (2004) 167–171.

[10] R.D. Sheth, S. Goldstein, Comparison of 1-84 and intact parathyroid hormone assay in pediatric dialysis patients, Pediatr. Nephrol 20 (2005) 977–981.
L.J. Sokoll, H. Drew, R. Udelsman, Intraoperative parathyroid hormone analysis: a study of 200 consecutive cases, Clin. Chem. 46 (2000) 1662–1668.

[11] S. Waller, D. Ridout, T. Cantor, L. Rees, Differences between "intact" PTH and 1-84 PTH assays in chronic renal failure and dialysis, Pediatr. Nephrol. 20 (2005) 197–199.

[12] H. Yamashita, T. Cantor, S. Uchino, S. Watanabe, T. Ogawa, T. Moriyama, et al., Sequential changes in plasma intact and whole parathyroid hormone levels during parathyroidectomy for secondary hyperparathyroidism, World J. Surg. 29 (2005) 169–173.

[13] R.C. Gensure, T.J. Gardella, H. Juppner, Parathyroid hormone and parathyroid hormone-related peptide, and their receptors, Biochem. Biophys. Res. Commun. 328 (2005) 666–678.

[14] T.O. Carpenter, CYP24A1 loss of function: clinical phenotype of monoallelic and biallelic mutations, J. Steroid Biochem. Mol. Biol. 173 (2017) 337–340.

[15] F. Manz, R. Kehrt, B. Lausen, A. Merkel, Urinary calcium excretion in healthy children and adolescents, Pediatr. Nephrol. 13 (9) (1999) 894–899.

[16] P.L. Riches, E. McRorie, W.D. Fraser, C. Determann, R. van't Hof, S.H. Ralston, Osteoporosis associated with neutralizing autoantibodies against osteoprotegerin, N. Engl. J. Med. 361 (15) (2009) 1459–1465.

[17] M.P. Whyte, S.E. Obrecht, P.M. Finnegan, J.L. Jones, M.N. Podgornik, W.H. McAlister, et al., Osteoprotegerin deficiency and juvenile Paget's disease, N. Engl. J. Med. 347 (3) (2002) 175–184.

[18] J. Griffiths, A. Vernocchi, E. Simoni, Transient hyperphosphatasemia of infancy and childhood. A study of serum alkaline phosphatase by electrofocusing techniques, Arch. Pathol. Lab. Med. 119 (9) (1995) 784–789.

[19] A.C. Looker, D.C. Bauer, C.H. Chesnut III, C.M. Gundberg, M.C. Hochberg, G. Klee, et al., Clinical use of biochemical markers of bone remodeling: current status and future directions, Osteoporos. Int. 11 (2000) 467–480.

[20] B.L. Langdahl, Is there a place for bone turnover markers in the management of osteoporosis? J. Bone Miner. Res. 33 (7) (2018) 1197–1198.

[21] A.A. Bahamat, M. Assidi, S.A. Lary, M.M. Almughamsi, A.A. Peer Zada, A. Chaudhary, et al., Use of array comparative genomic hybridization for the diagnosis of DiGeorge syndrome in Saudi Arabian population, Cytogenet. Genome Res. 154 (1) (2018) 20–29.

[22] E.M. Brown, Anti-parathyroid and anti-calcium sensing receptor antibodies in autoimmune hypoparathyroidism, Endocrinol. Metab. Clin. North Am. 38 (2) (2009) 437–445.

[23] N. Tomar, E. Kaushal, M. Das, N. Gupta, C. Betterle, R. Goswami, Prevalence and significance of NALP5 autoantibodies in patients with idiopathic hypoparathyroidism, J. Clin. Endocrinol. Metab. 97 (4) (2012) 1219–1226.

[24] M. Habibullah, J.A. Porter, N. Kluger, A. Ranki, K.J.E. Krohn, M. L. Brandi, et al., Calcium-sensing receptor autoantibodies in patients with autoimmune polyendocrine syndrome type 1: epitopes, specificity, functional affinity, IgG subclass, and effects on receptor activity, J. Immunol. 201 (11) (2018) 3175–3183.

[25] H.J. van der Vliet, Nieuwenhuis EE.IPEX as a result of mutations in FOXP3, Clin. Dev. Immunol. 2007 (2007) 89017.

[26] M. Alimohammadi, P. Björklund, A. Hallgren, N. Pöntynen, G. Szinnai, N. Shikama, et al., Autoimmune polyendocrine syndrome type 1 and NALP5, a parathyroid autoantigen, N. Engl. J. Med. 358 (10) (2008) 1018–1028.

[27] Y. Li, Y.H. Song, N. Rais, E. Connor, D. Schatz, A. Muir, et al., Autoantibodies to the extracellular domain of the calcium sensing receptor in patients with acquired hypoparathyroidism, J. Clin. Invest. 97 (4) (1996) 910–914.

[28] M.D.P. Larosa, R. Mackenzie, P. Burne, S. Garelli, S. Barollo, S. Masiero, et al., Assessment of autoantibodies to interferon-ω in patients with autoimmune polyendocrine syndrome type 1: using a new immunoprecipitation assay, Clin. Chem. Lab. Med. 55 (7) (2017) 1003–1012.

[29] A. Linglart, M.A. Levine, H. Jüppner, Pseudohypoparathyroidism, Endocrinol. Metab. Clin. North. Am. 47 (4) (2018) 865–888.

[30] G. Mantovani, M. Bastepe, D. Monk, L. de Sanctis, S. Thiele, A. Usardi, et al., Diagnosis and management of pseudohypoparathyroidism and related disorders: first International Consensus Statement, Nat. Rev. Endocrinol. 14 (8) (2018) 476–500.

[31] A.J. Williams, J.L. Wilkinson, W.H. Taylor, Pseudohypoparathyroidism. Variable manifestations within a family, Arch. Dis. Child. 52 (10) (1977) 798–800.

[32] J.G.H. Bruystens, J. Wu, A. Fortezzo, J. Del Rio, C. Nielsen, D.K. Blumenthal, et al., Structure of a PKA RIα recurrent acrodysostosis mutant explains defective cAMP-dependent activation, J. Mol. Biol. 428 (24 Pt B) (2016) 4890–4904.

[33] J. Hoppmann, J. Gesing, C. Silve, C. Leroy, A. Bertsche, F.W. Hirsch, et al., Phenotypic variability in a family with acrodysostosis type 2 caused by a novel PDE4D mutation affecting the serine target of protein kinase-A phosphorylation, J. Clin. Res. Pediatr. Endocrinol. 9 (4) (2017) 360–365.

[34] J.M. Gertner, S. Tomlinson, J. Gonzalez-Macias, Normocalcaemic pseudohypoparathyroidism with unusual phenotype, Arch. Dis. Child. 53 (4) (1978) 312–315.
[35] M. Christov, H. Jüppner, Phosphate homeostasis disorders, Best Pract. Res. Clin. Endocrinol. Metab. 32 (5) (2018) 685–706.
[36] Y. Xu, T. Hashizume, M.C. Shuhart, C.L. Davis, W.L. Nelson, T. Sakaki, et al., Intestinal and hepatic CYP3A4 catalyze hydroxylation of 1alpha,25-dihydroxyvitamin D(3): implications for drug-induced osteomalacia, Mol. Pharmacol. 69 (1) (2006) 56–65.
[37] A. Giustina, R.A. Adler, N. Binkley, R. Bouillon, P.R. Ebeling, M. Lazaretti-Castro, et al., Controversies in vitamin D: summary statement from an international conference, J. Clin. Endocrinol. Metab. (2018).
[38] N. Binkley, D. Krueger, C.S. Cowgill, L. Plum, E. Lake, K.E. Hansen, et al., Assay variation confounds the diagnosis of hypovitaminosis D: a call for standardization, J. Clin. Endocrinol. Metab. 89 (7) (2004) 3152–3157.
[39] M. Bedner, K.A. Lippa, S.S.-C. Tai, An assessment of 25-hydroxyvitamin D measurements in comparability studies conducted by the Vitamin D Metabolites Quality Assurance Program, Clin. Chim. Acta 426 (2013).
[40] R. Bouillon, G. Carmeliet, Vitamin D insufficiency: definition, diagnosis and management, Best Pract. Res. Clin. Endocrinol. Metab. 32 (5) (2018) 669–684.
[41] A.M. Willett, Vitamin D status and its relationship with parathyroid hormone and bone mineral status in older adolescents, Proc. Nutr. Soc. 64 (2) (2005) 193–203.
[42] A.C. Ross, J.E. Manson, S.A. Abrams, J.F. Aloia, P.M. Brannon, S.K. Clinton, et al., The 2011 report on dietary reference intakes for calcium and vitamin D from the Institute of Medicine: what clinicians need to know, J. Clin. Endocrinol. Metab. 96 (1) (2011) 53–58.
[43] F. Alshahrani, N. Aljohani, Vitamin D: deficiency, sufficiency and toxicity, Nutrients. 5 (9) (2013) 3605–3616.
[44] Institute of Medicine, Dietary reference intakes for calcium and vitamin D, The National Academies Press, Washington, DC, 2011.
[45] N. Charoenngam, A. Shirvani, M.F. Holick, Vitamin D for skeletal and non-skeletal health: what we should know, J. Clin. Orthop. Trauma. 10 (6) (2019) 1082–1093.
[46] 25-hydroxy vitamin D concentration reference interval. <https://www.mayocliniclabs.com/test-catalog/Clinical + and + Interpretive/83670> (accessed 02.18.19).
[47] 25-hydroxy vitamin D concentration reference interval. <https://testdirectory.questdiagnostics.com/test/test-detail/92888/questassured-25-hydroxyvitamin-d-d2-d3?p = r&q = 25-oh%20vitamin&cc = MASTER> (accessed 12.18.19).
[48] 25-hydroxy vitamin D concentration reference interval. <https://www.labcorp.com/tests/081950/vitamin-d-25-hydroxy> (accessed 12.18.19).
[49] 25-hydroxy vitamin D concentration reference interval. <http://ltd.aruplab.com/Tests/Pub/2002348> (accessed 12.29.19).
[50] K. Nitta, A. Yajima, K. Tsuchiya, Management of osteoporosis in chronic kidney disease, Intern. Med. 56 (24) (2017) 3271–3276.
[51] L.E. Silberstein, S. Naryshkin, J.J. Haddad, J.F. Strauss 3rd, Calcium homeostasis during therapeutic plasma exchange, Transfusion 26 (2) (1986) 151–155PMID:. Available from: 3006298.
[52] P. Makras, S.E. Papapoulos, Medical treatment of hypercalcaemia, Hormones 8 (2) (2009) 83–95.
[53] D.A. Anastasilakis, P. Makras, S.A. Polyzos, A.D. Anastasilakis, Asymptomatic and normocalcemic hyperparathyroidism, the silent attack: a combo-endocrinology overview, Hormones (2018).
[54] M. Pawlowska, N.E. Cusano, An overview of normocalcemic primary hyperparathyroidism, Curr. Opin. Endocrinol. Diabetes Obes. 22 (6) (2015) 413–421.
[55] A.E. Stephen, M. Mannstadt, R.A. Hodin, Indications for surgical management of hyperparathyroidism: a review, JAMA Surg. 152 (9) (2017) 878–882.
[56] L. Yuan, J. Liu, Y. Kan, J. Yang, X. Wang, The diagnostic value of 11C-methionine PET in hyperparathyroidism with negative 99mTc-MIBI SPECT: a meta-analysis, Acta Radiol. 58 (5) (2017) 558–564.
[57] W.A.M. Broos, M. Wondergem, R.J.J. Knol, F.M. van der Zant, Parathyroid imaging with 18F-fluorocholine PET/CT as a first-line imaging modality in primary hyperparathyroidism: a retrospective cohort study, EJNMMI Res. 9 (1) (2019) 72.
[58] A.B. Carter, P.J. Howanitz, Intraoperative testing for parathyroid hormone: a comprehensive review of the use of the assay and the relevant literature, Arch. Pathol. Lab. Med. 127 (2003) 1424–1442.
[59] J. Norman, D. Politz, 5,000 parathyroid operations without frozen section or PTH assays: measuring individual parathyroid gland hormone production in real time, Ann. Surg. Oncol. 16 (3) (2009) 656–666.
[60] D. Carneiro-Pla, Contemporary and practical uses of intraoperative parathyroid hormone monitoring, Endocr. Pract. 17 (Suppl 1) (2011) 44–53.
[61] G.L. Irvin 3rd, A.S. Molinari, C. Figueroa, D.M. Carneiro, Improved success rate in reoperative parathyroidectomy with intraoperative PTH assay, Ann. Surg. 229 (6) (1999) 874–878.
[62] L.M.1 Brunt, G.A. Sicard, Current status of parathyroid autotransplantation, Semin. Surg. Oncol. 6 (2) (1990) 115–121.
[63] J.A. Sloand, M.A. Shelly, Normalization of lithium-induced hypercalcemia and hyperparathyroidism with cinacalcet hydrochloride, Am. J. Kidney Dis. 48 (5) (2006) 832–837.
[64] J.Y. Lee, D.M. Shoback, Familial hypocalciuric hypercalcemia and related disorders, Best. Pract. Res. Clin. Endocrinol. Metab. 32 (5) (2018) 609–619.
[65] C.M. Gorvin, A. Rogers, B. Hastoy, A.I. Tarasov, M. Frost, S. Sposini, et al., AP2σ mutations impair calcium-sensing receptor trafficking and signaling, and show an endosomal pathway to spatially direct G-protein selectivity, Cell Rep. 22 (4) (2018) 1054–1066.
[66] R. Alrezk, F. Hannah-Shmouni, C.A. Stratakis, MEN4 and CDKN1B mutations: the latest of the MEN syndromes, Endocr. Relat. Cancer. 24 (10) (2017) T195–T208.
[67] V.K. Bahl, A. Sandhu, V. Mohan, S. Nolan, Multiple endocrine neoplasia and polyglandular autoimmune syndrome: a new association, Endocr. Pract. 4 (4) (1998) 208–212.
[68] A. Machens, H. Dralle, Advances in risk-oriented surgery for multiple endocrine neoplasia type 2, Endocr. Relat. Cancer 25 (2) (2018) T41–T52.

[69] F. Raue, K. Frank-Raue, Genotype-phenotype correlation in multiple endocrine neoplasia type 2, Clinics 67 (Suppl. 1) (2012) 69–75.
[70] R.V. Thakker, Multiple endocrine neoplasia type 1 (MEN1) and type 4 (MEN4), Mol. Cell. Endocrinol. 386 (1–2) (2014) 2–15.
[71] M. Iacobone, G. Masi, L. Barzon, A. Porzionato, V. Macchi, F.A. Ciarleglio, et al., Hyperparathyroidism-jaw tumor syndrome: a report of three large kindred, Langenbecks Arch. Surg. 394 (5) (2009) 817–825.
[72] K. Koikawa, Y. Okada, H. Mori, M. Kawaguchi, S. Uchino, Y. Tanaka, Hyperparathyroidism-jaw tumor syndrome confirmed by preoperative genetic testing, Intern. Med. 57 (6) (2018) 841–844.
[73] S.J. Marx, D.M. Lourenço Jr., Familial hyperparathyroidism - disorders of growth and secretion in hormone-secretory tissue, Horm. Metab. Res. 49 (11) (2017) 805–815.
[74] G. Coen, E. Bonucci, P. Ballanti, A. Balducci, S. Calabria, G.A. Nicolai, et al., PTH 1–84 and PTH "7–84" in the noninvasive diagnosis of renal bone disease, Am. J. Kidney Dis. 40 (2002) 348–354.
[75] H. Yamashita, P. Gao, T. Cantor, S. Noguchi, S. Uchino, S. Watanabe, et al., Comparison of parathyroid hormone levels from the intact and whole parathyroid hormone assays after parathyroidectomy for primary and secondary hyperparathyroidism, Surgery 135 (2004) 149–156.
[76] H. Yamashita, P. Gao, S. Noguchi, T. Cantor, S. Uchino, S. Watanabe, et al., Role of cyclase activating parathyroid hormone (1–84 PTH) measurements during parathyroid surgery: potential improvement of intraoperative PTH assay, Ann. Surg. 236 (2002) 105–111.
[77] G.D. Roodman, Osteoblast function in myeloma, Bone. 48 (1) (2011) 135–140.
[78] <https://www-ncbi-nlm-nih-gov.lp.hscl.ufl.edu/books/NBK278960/> (accessed 02.12.18).
[79] P.J. Tebben, R.J. Singh, R. Kumar, Vitamin D-mediated hypercalcemia: mechanisms, diagnosis, and treatment, Endocr. Rev. 37 (5) (2016) 521–547.
[80] H. Saito, H. Noda, P. Gatault, D. Bockenhauer, K.Y. Loke, O. Hiort, et al., Progression of mineral ion abnormalities in patients with Jansen metaphyseal chondrodysplasia, Clin. Endocrinol. Metab. 103 (7) (2018) 2660–2669.
[81] A. Kumar, V.K. Jain, M. Bharadwaj, R.K. Arya, Ollier disease: pathogenesis, diagnosis, and management, Orthopedics. 38 (6) (2015) e497–e506.
[82] I. Yunus, A. Fasih, Y. Wang, The use of procalcitonin in the determination of severity of sepsis, patient outcomes and infection characteristics, PLoS One. 13 (11) (2018) e0206527.
[83] M.S. Razzaque, The FGF23-Klotho axis: endocrine regulation of phosphate homeostasis, Nat. Rev. Endocrinol. 5 (11) (2009) 611–619.
[84] C. Holroyd, C. Cooper, E. Dennison, Epidemiology of osteoporosis, Best Pract. Res. Clin. Endocrinol. Metab. 22 (5) (2008) 671–685.
[85] M.C. Vlot, M. den Heijer, R.T. de Jongh, M.G. Vervloet, W.F. Lems, R. de Jonge, et al., Clinical utility of bone markers in various diseases, Bone. 114 (2018) 215–225.
[86] P. Glendenning, S.A.P. Chubb, S. Vasikaran, Clinical utility of bone turnover markers in the management of common metabolic bone diseases in adults, Clin. Chim. Acta 481 (2018) 161–170.

Self-assessment questions

1. Where is 25-hydroxyvitamin D normally converted into 1,25-dihydroxyvitamin D?
   - **a.** liver
   - **b.** intestine
   - **c.** parathyroid gland
   - **d.** C cell
   - **e.** kidney tubule
2. What is the etiology of hypercalcemia in sarcoidosis?
   - **a.** excess PTH acting on the renal tubule to increase Ca and phosphate absorption
   - **b.** depressed calcitonin production by the C cells
   - **c.** increased conversion of 25-hydroxyvitamin D to 1,25-dihydroxyvitamin D by macrophages
   - **d.** increased vitamin D action on the intestine
   - **e.** increased albumin concentrations
3. In biliary atresia, rickets results from:
   - **a.** decreased bile flow and decreased vitamin D absorption
   - **b.** decreased conversion of vitamin D into 25-hydroxyvitamin D
   - **c.** both
   - **d.** neither
4. 4. The pathogenesis of renal osteodystrophy involves all of the following processes except:
   - **a.** phosphate retention, depressed calcium, and increased PTH.
   - **b.** decreased conversion of 25-hydroxyvitamin D into 1,25-dihydroxyvitamin D.
   - **c.** hyperparathyroidism.
   - **d.** resistance to the effects of PTH on bone.
   - **e.** decreased intestinal calcium absorption.
5. A 5-year-old girl's parents bring their adopted daughter to the pediatrician because she has suffered periodic episodes of what the parents call "stiffening." The intellectual development of this girl is significantly delayed. As well, she has short stature. Physical examination reveals a "heart-shaped" face, mild obesity, delayed development, and short fourth metacarpals. The Chvostek sign and Trousseau sign are elicited on physical examination. When parathyroid hormone was isolated from this child by immunoaffinity chromatography, purified, and injected into a normal mouse, a transient hyperphosphaturia was produced. The following laboratory studies are obtained on the child (see below). What is the most likely diagnosis?

	Result	Reference interval
Ca^{++}	6.5 mg/dL	8.4–10.4
PO_4^{---}	6.1 mg/dL	2.5–4.5
PTH	85 pg/mL	10–65
Albumin	4.0 g/dL	3.5–5
Cr	0.5 mg/dL	0.4–1.2
BUN	10 mg/dL	5–20

   - **a.** PTH/PTHrP receptor loss-of-function mutation
   - **b.** hypoparathyroidism, etiology not otherwise specified
   - **c.** loss-of-function mutation in the parathyroid calcium receptor
   - **d.** renal failure with secondary hyperparathyroidism
   - **e.** PTH loss-of-function mutation

Answers

1. e
2. c
3. b
4. d
5. a

Chapter 46

Vitamins: functions and assessment of status through laboratory testing

Emile Levy[1,2] and Edgard Delvin[1,3]

[1]*Division of Gastroenterology, Hepatology and Nutrition, CHU Sainte-Justine Research Centre, Montreal, QC, Canada,* [2]*Department of Nutrition, Faculty of Medicine, University of Montreal, Montreal, QC, Canada,* [3]*Department of Biochemistry, Faculty of Medicine, University of Montreal, Montreal, QC, Canada*

Learning objectives

After reviewing this chapter, the reader should be able to:

- Recognize clinically important lipid- and water-soluble vitamins.
- Describe situations in which their assessment status is necessary to prevent deterioration in health status of individuals and/or populations.
- Apply vitamin testing to situations requiring the investigation of their status utilizing methods available to clinical and research laboratories.
- Discuss the methodological approaches used to measure the wide variety of lipid- and water-soluble vitamins.

Introduction

According to the National Institute of Health (NIH; Bethesda, MD, United States), "the science of nutrition interprets the interaction of nutrients and other substances in food in relation to maintenance, growth, reproduction, health, and disease of an organism [1]. It covers food intake, absorption, assimilation, biosynthesis, catabolism, and excretion." Nutritional biochemistry per se encompasses the study of the chemistry and biology of nutrients. Societies such as the American Society for Nutrition (www.nutrition.org), the Academy of Nutrition and Dietetics (www.eatright.org), and the European Society for Clinical Nutrition and Metabolism (http://www.espen.org) publish either in print or online a vast repertoire of literature, making this science and the metabolism and measurement of nutrients more accessible both to health professionals and lay public. Furthermore, government-sponsored panels of nutrition leaders produce consensus statements and materials to guide nutrition thought that could be consulted in internationally well-known nutrition-related journals. As the micronutrient market, *claiming beneficial effects in preventing deficiencies or enhancing physical and intellectual capacities*, is a multi-billion dollar business, it must be underlined that all online materials may not be authoritative, and can even be deceptive or even fraudulent. Therefore it is the obligation of the professionals in the fields to guide consumers in distinguishing the oat from the chaff. This requires the development of tools for sturdy and reproducible nutritional status assessment schemes.

Nutrients are compounds or elements that the body must use for essential metabolic processes including growth. They are broadly classified into two categories: the macronutrients and the micronutrients. Macronutrients, which include protein (Chapter 22, Proteins: analysis and interpretation of serum, urine, and cerebrospinal fluid), carbohydrates (Chapter 33, Evaluation of exocrine pancreatic function), and lipids and lipoproteins (Chapter 27, Diagnostic body fluid testing), consumed in amounts of >1 g/day, are sources of energy required for metabolism and for synthesis of cell structural and functional components. Micronutrients (consumed in amounts of <1 mg/day) include vitamins, minerals, trace elements, and ultra-trace elements. Their physiologic requirements change according to statuses related, among others, to growth, pregnancy, lactation, activity, aging, and disease. This chapter covers the lipid- and water-soluble vitamins.

Nota bene.[1]

1. *Nota bene*: Reference intervals *provided* in the text for each vitamin are obtained from studies that include variable number of subjects; some having fewer than 100 participants, and some based on population studies. Hence, the values cited should be interpreted with caution and are mainly suggestive.

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00046-6

Vitamins

Three criteria define the term vitamin: (1) vitamin deficiency produces symptoms or a disease that can be corrected by returning the micronutrient to adequate levels; (2) vitamins have organic structures; and (3) vitamins are required in small amounts and are normally not synthesized by the body. Classically, vitamins have been classified as lipid- and water-soluble compounds, a characteristic that dictates their cellular uptake and metabolism.

Fat-soluble vitamins

Intestinal uptake and transport

The lipid-soluble vitamins A, D, E, and K must associate with mixed micelles, consisting of phospholipids, lysophospholipids, monoacylglycerol, cholesterol, and bile salts, or with vesicles in order to be absorbed at the enterocyte apical brush border and transported to the basolateral membrane to be distributed to the various target organs in order to exert their biological effects. Although initial studies led to the conclusion that fat-soluble vitamins were absorbed by passive diffusion, more recent evidence, reviewed comprehensively [2–4], shows that, while passive diffusion occurs at pharmacological doses, absorption at physiological or dietary doses requires protein carriers, such as Niemann-Pick C1-like 1, cluster determinant 36, ATP-binding cassette subfamily A member 1, and scavenger receptor class B type 1, all involved in lipid transport and trafficking. Fig. 46.1 schematizes this process. It is well-known that congenital lipid malabsorption diseases, such as chylomicron retention disease, familial abetalipoproteinemia, or familial hypobetalipoproteinemia, involving genes related to intestinal lipoprotein assembly and secretion, and also to lipid-soluble uptake, lead to severe conditions such as retinopathy, myopathy, neuropathy, cardiomyopathy hepatic steatosis, steatorrhea, and acanthosis (Table 46.1) [5]. It can be argued that these conditions could be due, at least in part, to deficiency in fat-soluble vitamins. The presence of cholesterol transporters, facilitating also the transport of lipid-soluble vitamins, could lessen the severity of the deficiency. If deficiencies impact on health status, so do excesses. Table 46.2 summarizes the effect of lipid-soluble vitamin deficiencies and excesses on the health status in the general population. Thorough evaluation of the vitamin statuses is thus warranted in those conditions.

Vitamin A

According to the IUPAC-IUB Joint Commission on Biochemical Nomenclature recommendations of 1981, the term vitamin A is a generic descriptor for retinoids, exhibiting the biological activity of *all-trans* retinol [6]. To exert its role as activator of nuclear transcription factors, *all-trans* retinol must be converted to *all-trans* retinoic acid by ubiquitous cytochrome P450s [7,8]. Preformed vitamin A is derived from animal sources, particularly fish liver oils and liver. Plants, principally yellow-pigmented vegetables and fruits, provide α-carotene, β-carotene, and β-cryptoxanthin that must be converted, in the intestine and liver, to vitamin A-active compounds. Some retinoids, such as 11-*cis* retinaldehyde, have very specific functions. This retinoid covalently binds with, the rod outer segment photoreceptor protein, rhodopsin through a protonated Schiff-base link. Photoactivation of this complex requires the isomerization of the bound 11-*cis* retinal moiety to the all-*trans* retinal, which, in turn, induces major structural changes of the receptor [9].

Vitamin A deficiency is a significant worldwide problem for public health, as it contributes significantly to the global disease burden that includes xerophthalmia and decreased resistance to infections, both of which are preventable [10,11]. *Vitamin A may however be a double-edged sword, having a higher potential for toxicity than water-soluble vitamins*. Indeed, intakes marginally above the recommended dietary intake have been associated with teratogenic effects [12]. In other clinical settings, neuromuscular adverse effects have been reported for

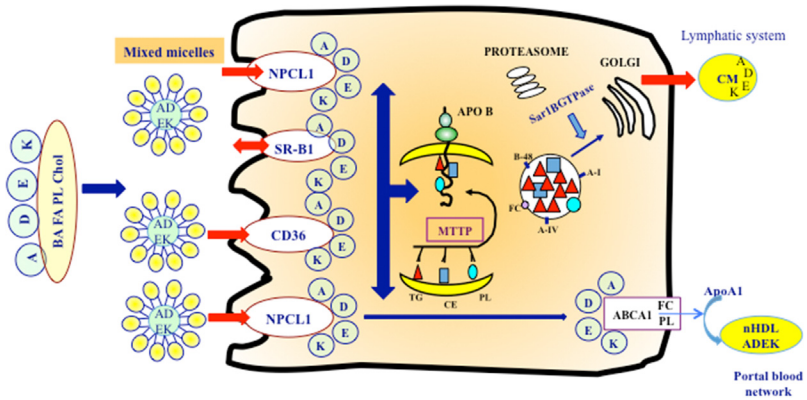


FIGURE 46.1 Simplified scheme of the uptake, transport, and secretion of vitamins A, D, E, and K by the enterocyte. At physiological doses, vitamins A, D, E, and K are captured from the mixed micelles by apical membrane transporters and transported across the cytosol to be secreted into the lymphatic system associated with chylomicrons or in the portal blood network associated with nascent HDL particles. A, D, E, and K: vitamins A, D, E, and K, respectively. *ABCA1*, ATP-binding cassette subfamily A member; *CD36*, cluster determinant 36; MTTP, microsomal triglyceride transfer protein; *NPC1-L1*, Niemann-Pick C1-like 1; *SR-B1*, scavenger receptor class B type 1.

TABLE 46.1 Consequences of lipid malabsorption.

Disorders	CRD	FHBL + / +	FHBL + / −	FABL
Retinopathy	+	+ + +	−	+ + +
Myopathy	+	+ + +	+/ −	+ + +
Neuropathy	+	+ + +	+/ −	+ + +
Cardiomyopathy	+/ −	+ +	−	+ +
Hepatic steatosis	+	+ +	+ +	+ +
Steatorrhea	+ +	+ + +	+/ −	+ + +
Acanthosis	−	+	+/-	+

Notes: (+) indicates severity of disease.
CRD: Chylomicron retention disease; *FABL*: familial abetalipoproteinemia; *FHBL*: familial hypobetalipoproteinemia.
Source: Adapted from N. Peretti, A. Sassolas, C.C. Roy, C. Deslandres, M. Charcosset, J. Castagnetti, et al., Guidelines for the diagnosis and management of chylomicron retention disease based on a review of the literature and the experience of two centers, Orphanet J. Rare Dis. 5 (2010) 24.

TABLE 46.2 Selected illnesses related to improper intake of lipid-soluble vitamins.

Vitamins	Deficiency	Excess
Vitamin A	Xerophthalmia, night blindness, and immunity disorders	Hypervitaminosis A (teratogenicity, myalgia, polyneuropathies, liver fibrosis, cholestasis, and hypercalcemia)
Vitamin D	Rickets, osteomalacia, and osteoporosis	Hypercalcemia and nephrocalcinosis
Vitamin E	Ataxia	Coagulopathies
Vitamin K	Coagulopathy and hemorrhagic disease of the newborn	Rare condition. High doses of vitamin K_3 (menadione) may produce red blood cell fragility

isotretinoin and acitretin, two commonly oral potent retinoids used in systemic dermatotherapy that have led to the development of neuromuscular algorithm for appropriate monitoring of patients treated with these two compounds [13]. The use of over-the-counter supplements is also a source of concern, as a case of vitamin A intoxication translating into deranged liver function tests consistent with a cholestatic process has been reported [14]. Hypercalcemia leading to nephrocalcinosis and renal impairment is yet another potential adverse effect of hypervitaminosis A as described by Safi et al. [15] in a 4-year-old cystic fibrosis patient. Although anecdotal, these case reports call physicians and clinical laboratory to vigilance when assessing nutritional status of patients.

Biochemical assessment of vitamin A status

Dietary preformed vitamin A, resulting from the obligatory hydrolysis in the intestinal lumen of long-chain fatty acid retinol esters, and retinol resulting from the cleavage of β-carotene are taken up by the mucosal cells, reesterified and incorporated with other neutral lipid esters, assembled into chylomicrons and secreted into the lymphatic system, and ultimately delivered to the liver where it is stored mainly as retinyl palmitate [16]. Vitamin A from the liver, providing the needs of the body, is stochiometrically bound to and transported by the retinol-binding protein (RBP) to the different target tissues [8]. When considering central clinical laboratories, the measurements of serum retinol and RBP concentrations are the standard methods to assess the vitamin A status. For specialized laboratories or for research purposes, relative dose response, modified relative dose response, and the deuterated retinol isotope dilution tests, reviewed by Tanumihardjo et al. [17], are also available.

A number of methods for the measurement of serum retinol have been developed in the last 40 years. The early methods involved fluorometric measurement with or without extraction and chromatography [18]. They were later replaced by high-performance liquid chromatography (HPLC) with ultraviolet (UV) detection systems. Methods for the measurement of retinol standardized based on the use of Standard Reference Material (SRM) have been reported [19,20]. However, at the present time, there is no international reference method available, although the Micronutrients Measurement Quality Assurance Program

(M2QAP) of the National Institute of Standards and Technology (NIST; Gaithersburg, MD, United States) is available [21]. The recent development of accessible mass spectrometric methods offers the advantage metabolic profiling of the different circulating retinoids and hence enables to measure retinoic acid, the active component of the vitamin [22]. Serum RBP has also been used as a surrogate for serum retinol concentrations. Although RBP has the advantage of being more stable than retinol, this protein exists in multiple forms. Immunoassays have so far not been standardized, and there are no international methods available with which a standardization or harmonization could be performed. Blood plasma retinol concentrations are typically 1−2 μM and are not affected by daily intake [8].

Vitamin D

The widespread term vitamin D makes no distinction between vitamin D_3 (cholecalciferol) of human or animal origin, and vitamin D_2 (ergocalciferol) of plant origin. The term "vitamin," that is, a "vital" product that the organism cannot produce is inappropriate for vitamin D, as the skin can synthesize vitamin D_3 from 7-dehydrocholesterol under the action of UVB radiation. The role of vitamin D in bone health is well established, and its suspected role in other systems has been addressed [23,24]. Despite major national public health campaigns since the late 1960s, Vitamin D insufficiency remains a common problem, particularly in high latitudes [25−27], leading to increased incidence of rickets, osteomalacia, and osteoporosis [28,29]. Vitamin D toxicity, also known as hypervitaminosis D, remains anecdotal despite an increased risk with vitamin D supplementation practices [30]. However, advocacy of high vitamin D intakes in the order of 10,000 IU/d warrants caution, and rather guidelines published by the Institute of Medicine should be applied [31,32].

Vitamin D intoxication is a serious condition. It presents with hypercalcemia, poor appetite, weight loss, abdominal pain, vomiting, constipation, polyuria, and polydipsia, and in severe situations, life-threatening dehydration. In some cases, the concentration of calcium phosphate salts in the glomerular filtrate may exceed its solubility threshold, resulting in nephrocalcinosis, and in extreme cases to renal tubular acidosis and insufficiency [33−36].

Biochemical assessment of vitamin D status

Vitamin D nutritional status is evaluated by the measurement of serum 25-hydroxyvitamin D (25OHD), and not by the parent vitamin D or the hormonal form 1α,25$(OH)_2$D. In cases of vitamin D deficiency, serum 1α,25$(OH)_2$D may be normal, low or high depending on the secretion of parathyroid hormone (PTH) and FGF-23 [37,38]. Lips [29,39] suggested defining vitamin D deficiency as that serum 25OHD concentration below which there are detrimental effects on health, particularly on bone because of hyperparathyroidism. The relationship between circulating 25OHD and PTH concentrations is of interest, as it theoretically allows defining "functional vitamin D deficiency." The results from the locally weighted scattered plot smoothing (Loess) model, as shown in Fig. 46.2, exemplifies that this relationship may not be as clear as wished with a natural breakpoint 25OHD concentration ranging from 40 to 50 nmol/L [41]. Furthermore, Steingrimsdottir et al. [42] reported a calcium intake−optimal 25OHD interaction for an effect on circulating PTH levels, which was confirmed by Aloia et al. [41]. Patel et al. [43] recently confirmed the calcium dependency of the 25OHD−PTH relationship in healthy adolescents. Fig. 46.3 adapted from their study shows that circulating intact PTH levels are lower in participants with calcium intakes of >520 mg/day independently of serum 25OHD concentration. Furthermore, free-hand drafting of asymptotes for the respective axes suggests that the threshold for defining "functional" vitamin D deficiency (inflexion point of increased PTH levels) is situated at ≈12 ng/mL (30 nmol/L).

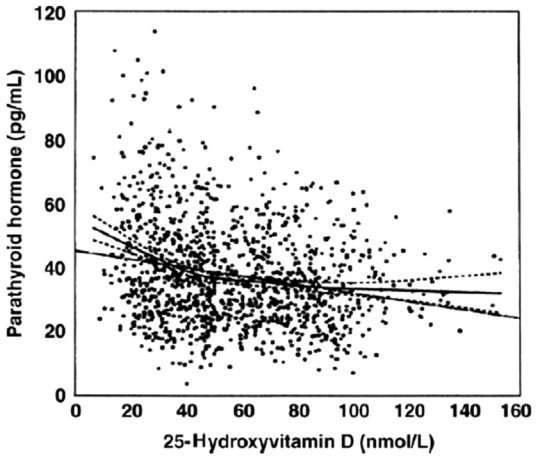


FIGURE 46.2 Relationship between circulating parathyroid hormone and serum 25-hydroxyvitamin D concentrations. Loess plot (solid line) and 95% confidence interval of the Loess plot (dotted lines) of parathyroid hormone as a function of 25-hydroxyvitamin D ($n = 124$). The linear regression line (dashed line) is outside the 95% confidence interval for 25-hydroxyvitamin D concentrations of <30 or >110 nmol/L. *Reproduced with permission from J.F. Aloia, S.A. Talwar, S. Pollack, M. Feuerman and K.K. Yeh, Optimal vitamin D status and serum parathyroid hormone concentrations in African American women,* Am J Clin Nutr. ***84*** *(3), 2006, 602−609.*

The observed widespread differences in circulating 25OHD concentrations across epidemiological studies restrain from concluding on the circulating 25OHD concentrations required for optimal health status, and confuse

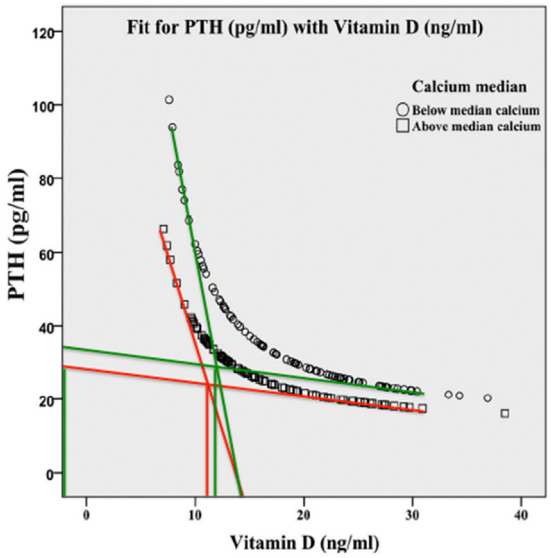


FIGURE 46.3 Relationship between plasma parathyroid hormone and 25-hydroxyvitamin D as a function of calcium intake. Red lines: above median calcium intake. Green lines: below median calcium intake. The respective intercepts represent the estimate of the serum 25-hydroxyvitamin D concentrations for functional vitamin D deficiency. At high calcium intake, the serum 25-hydroxyvitamin D threshold is lower than that at low calcium intake. *Adapted with permission from P. Patel, M.Z. Mughal, P. Patel, B. Yagnik, N. Kajale, R. Mandlik, et al., Dietary calcium intake influences the relationship between serum 25-hydroxyvitamin D3 (25OHD) concentration and parathyroid hormone (PTH) concentration, Arch. Dis. Child. 101 (4) (2016) 316–319.*

the efforts in developing international evidence-based public health guidelines. Despite recent technological advances, the measurement of circulating 25OHD remains a challenge, and many important issues have still to be resolved to obtain an accurate measure of serum 25OHD concentration. To solve this issue, the consortium composed of the NIH Office of Dietary Supplements, the Center for Disease Control National Center for Environmental Health, the NIST, and Ghent University (Ghent, Belgium) have established the Vitamin D Standardization Program having as mandate the promotion of methodological consistency for the measurement of 25OHD [44]. This consortium is thus advocating, based on the recommendations of Stockl et al. [45], an imprecision (CV) of $\leq 10\%$ and a bias of $\leq 5\%$ as current goals for the analytical performance of vitamin D assays in routine clinical laboratories [46]. The impact of this standardization progress is exemplified in the recent report of the Canadian Health Measure Survey, showing that prevalence of vitamin D insufficiency (20 ng/mL or 50 nmol/L) in the Canadian population, based on serum 25OHD concentrations observed after the standardization process, increased from 29% to 36.8% [47]. Le Goff et al. [48] reviewed the main elements and the difficulties of the automated and semiautomated methods for 25OHD measurement, from sample preparation to the analytical phase, as well as those related to mass spectrometry, and emphasized the need for standardization to better define the clinical decision thresholds of vitamin D nutritional status [49–51]. Mass spectrometry is now considered the state-of-the-art method for accurately measuring vitamin D metabolites. Several methods have been published in the recent years [52–55].

As mentioned in the "Introduction" section, there is mounting evidence indicating that the measurement of serum vitamin D binding protein, and *in fine* the definition of its polymorphisms, should be part of the vitamin D nutritional assessment workup. The Endocrine Society defines vitamin D deficiency as serum 25OHD values of <50 nM, and insufficiency as values between 52.5 and 72.5 nM [56].

Vitamin E

The generic term vitamin E represents a family of eight naturally occurring complex compounds that include α-, β-, γ-, and δ-tocopherols and tocotrienols, differing in the position and number of methyl groups on their benzene rings C5, C7, and C8 of the chromanol structure. Tocopherol differs from tocotrienols by having a trimethyltridecyl residue side chain with chiral carbons at C4', C8', and C12' rather than a phytyl side chain unsaturated at C4, C7, and C11 [57–59]. The natural α-tocopherol stereoisomer is (2'R, 4'R, 8'R)-α-tocopherol, which is the one preferentially incorporated into liver lipoproteins for peripheral export [58] and recognized by the α-tocopherol transport protein, responsible for maintaining circulating α-tocopherol concentrations [60,61].

Human vitamin E deficiency symptoms mainly include progressive peripheral neurologic disorders, which, if not treated, may ultimately result in ataxia [62,63] and cardiomyopathies [64]. The disparity of biomarkers and cutoffs used in the definition of vitamin E deficiency prevent an accurate measure of the prevalence of its dietary inadequacy. It is, however, accepted that symptomatic nutritional vitamin E deficiency is rare in the general population, although there are groups at risk particularly in developing countries [65–67]. Vitamin E is moderately toxic compared with vitamin A or D [36,68]. However, when used at very high doses, it can potentially lead to harmful effects by antagonizing leukocyte and platelet 5-lipoxygenase activity, resulting in inadequate thromboxane and leukotriene synthesis, decrease in blood coagulating capacity, and disruption of granulocyte and phagocyte antiinfective function [69]. Vitamin E has been reported to interact with vitamin K metabolism particularly in individuals at risk for vitamin K nutritional or iatrogenic deficiency [70]. The mechanisms underlying this deleterious interaction elude

explanations, although competitive inhibition of the vitamin K metabolism has been suggested [70].

Biochemical assessment of vitamin E status

There are a variety of published methods available for the measurement of vitamin E. The Emmerie–Engel chemical method based on the oxidation of total tocopherols to tocopherol quinine by $FeCl_3$ and the reaction of the resulting $FeCl_2$ with $\alpha-\alpha$-dipyridyl to a red color, and the functional test based on the red blood cell (RBC) hemolysis index have been used for decades, but are now considered outdated. The current recommended methods are HPLC-based and involve an initial organic extraction step, followed by the separation and either using UV (292 nm) or fluorescent (excitation wavelength of 205 nm and with an emission filter of 340 nm) detection systems. Vitamin E is often measured concurrently with vitamin A [19,20,71]. Hinchliffe et al. [72] recently described a method allowing the simultaneous quantification of vitamins A and E by liquid chromatography–tandem mass spectrometry. As for vitamin A, standardization of vitamin E is possible through the M2QAP at the NIST, and no reference-standardized method has yet been accepted. Proposed reference values for vitamin E lie between 14 and 40 μM, while values of <21 μM can be considered unhealthy [73].

Vitamin K

The term vitamin K represents a collection of homologous fat-soluble compounds derived from the core molecule 2-methyl-1,4-naphthoquinone. The naturally occurring phylloquinone (2-methyl-3-phytyl-1,4-naphthoquinone), termed vitamin K_1, the major dietary form of vitamin K, mainly originates from plants. Its absorption depends largely on protein-mediated mechanisms, shared with other fat-soluble vitamins [74]. Krzyzanowska et al. [75] recently reported a higher prevalence and severity of vitamin K deficiency in cystic fibrotic patients for which the presence of the F508del CFTR mutation constituted an independent risk factor. This observation implied that this transmembrane conductance regulator is involved directly or indirectly in vitamin K transport.

Menaquinones, collectively known as vitamin K_2, are a large series of compounds containing an unsaturated side chain with differing numbers of isoprenyl units linked to the C3 of the 2-methyl-1,4-naphthoquinone nucleus. They are produced by the normal jejunum and ileum Gram (+) flora and represent a substantial portion of the daily requirements [76]. Contrasting with the other fat-soluble vitamins, vitamin K has a rapid turnover, resulting in comparatively low tissue stocks [40].

The pleiotropic roles of vitamin K are mostly channeled through a common pathway involving the γ-carboxylation of the glutamyl side-chain residue of specific protein factors, thereby known as GLA-proteins. These GLA-proteins are involved in diverse biological systems, such as blood clotting (prothrombin), bone formation (osteocalcin), mobilization, and deposition of calcium in bone hydroxyapatite (GLA-bone matrix protein), and kidney GLA-proteins involved in the tubular calcium reabsorption and solubilization of calcium salts in urine [77,78]. Vitamin K-dependent proteins are also thought to have roles in cell signaling and brain lipid metabolism [79]. Hence, vitamin K deficiency is likely to affect those systems to varying degrees depending upon the level of deficiency. However, impairment of the coagulation pathway is certainly the most overt symptom [77]. In a search of the NIH PubMed database using "vitamin K toxicity" with case reports as filter, no toxicity case emerged.

Biochemical assessment of vitamin K status

The initial measurements of vitamin K depended upon bioassays based on the reduction of prothrombin time (PT) in anticoagulant stressed chicken [80]. PT and the international normalized ratio are the main tests currently used in clinical homeostasis laboratories to monitor vitamin K antagonists and also to indirectly assess vitamin K status [81]. Reversed-phase HPLC methods with postcolumn reduction and fluorescent detection have been described and are the usual biochemical method to measure serum or plasma phylloquinone [82,83]. However, more recently available mass spectrometry-based methods allow the measurement and metabolic profiling of phylloquinone [84,85]. The Mayo Clinic reports normal values for vitamin K_1 in serum to be from 0.10 to 2.20 μg/L [86].

Water-soluble vitamins

Transport and absorption

Despite their different structures, water-soluble vitamins are central to various cellular functions where they are essential for normal cellular functions, growth, and development. Their deficiency causes diverse clinical entities spanning from growth retardation to neurological syndromes. Table 46.3 summarizes the conditions related to improper intake of water-soluble vitamins. Although water-soluble vitamins are obtained from dietary sources, the bacterial source via the gut microbiota should not be underestimated. Whereas vitamins obtained from the diet are mainly absorbed in the jejunum and the ileum, the colon absorbs those produced by the microbiota. They are, however, all transported by carrier-mediated mechanisms. An exception

TABLE 46.3 Selected illnesses related to improper intakes of water-soluble vitamins.

Vitamins	Deficiency	Excess
Vitamin B_1 (thiamin)	Subclinical deficiency include headache, tiredness, anorexia, and muscle wasting. Frank deficiency causes beriberi, Wernicke's encephalopathy, and Korsakoff psychosis.	Rare and usually follow long-term parenteral nutrition. High doses cause headaches, nausea, insomnia, and tachycardia.
Vitamin B_2 (riboflavin)	Dermatitis, photophobia, itching, glossitis, dizziness, insomnia, muscle weakness, and anemia associated with reticulocytopenia.	No toxic or adverse reactions were reported.
Vitamin B_3 (niacin)	Pellagra, dizziness, vomiting, constipation or diarrhea, neurological disorders, fatigue, memory loss, and visual deficiency.	Stems essentially from its use in treating hypercholesterolemia. Symptoms include flushing, itching, nausea, and gastrointestinal disorders. Few cases of liver dysfunction following high-dose nicotinamide treatment.
Vitamin B_5 (pantothenic acid)	Extremely rare. Provoked experimentally. Fatigue, apathy, sleep disturbance, gastrointestinal disorders, numbness, paresthesia, and staggering gait.	No toxic or adverse reactions were reported.
Vitamin B_6 (pyridoxal phosphate)	Extremely rare in humans. Children fed overheated milk display weakness, irritability, weight loss, and insomnia. Inherited pyridoxine dependency translates into neonatal seizures. Large doses of pyridoxine are required to avoid severe mental retardation.	Reports of paresthesia, somnolence, stumbling gait, lack of muscle coordination, and sensory loss.
Vitamin B_7 (biotin)	Desquamating dermatitis, skin rash, hair loss, and ataxia, conjunctivitis. Deficient infants show hypotonia, lethargy, developmental delay, and withdrawn behavior. Inherited biotinidase deficiency creates the similar disorders.	No toxic or adverse reactions were reported.
Vitamin B_9 (folic acid)	Megaloblastic and macrocytic anemia, and moderate hyperhomocysteinemia. In utero developmental defects: spina bifida. Decreased since folate supplementation.	No toxic or adverse reactions were reported.
Vitamin B_{12} (cobalamin)	Megaloblastic and macrocytic anemia. Neurological disorders and moderate hyperhomocysteinemia.	No toxic or adverse reactions were reported. Anecdotal cases of allergic reactions.
Vitamin C (ascorbic acid)	Scurvy, fatigue, muscle weakness, aching joints, and advanced deficiency lead to anemia, gum bleeding, and delayed wound healing.	Gastrointestinal disorders, metabolic acidosis, renal acidosis (megadoses), and renal lithiasis.
Choline*	Rare occurrence. hepatocarcinoma and hepatosteatosis.	Body fishy odor, gastrointestinal disorders, anorexia, and mild cholinergic toxicity.

Based on *Standing Committee on the Scientific Evaluation of Dietary Reference Intakes and Its Panel on Folate OV, and Choline and Subcommittee on Upper Reference Levels of Nutrients, Food and Nutrition Board, Institute of Medicine, Pantothenic Acid. Dietary Reference Intakes for Thiamine, Riboflavin, Niacin, Vitamin B6, Folate, Vitamin B12, Pantothenic Acid, Biotin, and Choline, The National Academies Press, Washington, DC, 1998.

being made of niacin of limited endogenous synthesis; water-soluble vitamins must be obtained from either the diet or the colonic microbiota; hence, the body depends on their intestinal absorption. The mechanisms regulating their luminal absorption, cellular transport, and basolateral secretion are subjects of excellent reviews [88–91]. Table 46.4 lists the source and known carriers for the various water-soluble vitamins.

The B vitamin complex

The water-soluble vitamin B complex consists of eight vitamins all involved in cell function and survival. They are thiamine (vitamin B_1 and vitamin F), riboflavin (vitamin B_2 and vitamin G), niacin (vitamin B_3, vitamin P, or vitamin PP), pantothenic acid (vitamin B_5), pyridoxine and pyridoxamine (vitamin B_6), biotin (vitamin B_7, vitamin H, and Vitamin BW), folic acid (vitamin B_9), and cyanocobalamin (CNCbl; vitamin B_{12}). The magnitude of body stores varies for each vitamin; hence, the impact of dietary restriction changes, for example, from 4 to 10 days for thiamine to 3–5 years for vitamin B_{12} [92]. Water-soluble vitamin deficiencies are observed in a number of conditions, such as chronic renal insufficiency patients, in whom peripheral neurological disorders and anemia, documented in the early 1990s, have partly been corrected by vitamin supplements [93,94]. Whereas the small intestine and the pancreas absorb dietary and

TABLE 46.4 Sources and transporters of water-soluble vitamins.

Vitamin	Source	Transporter	Associated genes
Thiamine	Dietary	THTR1/THTR2	*SLC19A2/SLC19A3*
Riboflavin	Dietary	RFVT1/RFVT2	*SLC52A1/SLC52A2*
Niacin	Dietary/microbial	OAT10?	*SLC22A13?*
Pantothenic acid	Dietary/microbial?	SMVT	*SLC5A6*
Pyridoxine/pyridoxamine	Dietary/microbial	?	*?*
Biotin	Dietary/microbial	SMVT	*SLC5A6*
Folic acid	Dietary/microbial	FOLT/PCFT/HCP1/FOLR1	*SLC19A1/SLC46A1*
Cyanocobalamin	Dietary/microbial	Cubam (cubilin/amnionless)	*?*
Choline	Dietary/endogenous	OCT1/CTL?	*SLC22/SLC44?*
Ascorbic acid	Dietary	SVCT1/SVCT2	*SLC23A1/SLC23A2*

Notes: Cubam is composed of two proteins, cubilin (intrinsic factor-cobalamin receptor) and amnionless (amnion-associated transmembrane protein). Cubilin harbors ligand-binding capabilities, while amnionless provides membrane anchorage and potential endocytic capacity.
CTL, Choline transporter; *CVCT1/SVCT2*, specific vitamin C transporter, members 1 and 2; *FOLR1*, folate receptor 1; *FOLT*, folate transporter; *HCP1*, heme carrier protein 1; *OAT*, organic anion transporter; *OCT*, organic cation transporter; *PCFT*, proton-coupled folate transporter; *RFVT1/RFVT2*, riboflavin transporter, members 1 and 2; *SLC*, solute carrier; *SMVT*, Na^+-dependent specific vitamin carrier system; *THTR1/THTR2*, thiamine transporter, members 1 and 2.

supplement water-soluble vitamins, those produced by the microbiota are absorbed in the colon. The mechanisms involved in each process have recently been reviewed [90].

Thiamine

Thiamine, with the IUPAC name of 3-((4-amino-2-methyl-5-pyrimidinyl)methyl)-5-(2-hydroxyethyl)-4-methylthiazolium chloride, is synthesized by bacteria, fungi, and plants. It is present in the rice grain envelope. It is for that reason that, in the late 19th century endemic beriberi, manifesting as cardiovascular disease, central nervous system disorder (Wernicke-Korsakoff encephalopathy syndrome), or peripheral neuropathy, was described in the Dutch Indonesian people who, contrary to the indigenous people, were consuming exclusively polished rice. Interestingly, Chinese medical literature described the symptoms of beriberi as early as of the 10th century [95]. Thiamine plays key roles as a cofactor in several biochemical pathways, such as the respiratory chain in oxidative metabolism, the regulation of carbohydrate metabolic flux through α-ketoglutarate dehydrogenase, a highly regulated enzyme of the Krebs cycle, the biosynthesis of the neurotransmitters acetylcholine (ACh) and γ-aminobutyric acid, and myelin production. Oxidative-dependent organs such as the nervous system and the heart are particularly sensitive to thiamine deficiency that results in mitochondrial uncoupling [96]. Activation of thiamine requires sequential phosphorylation steps to thiamine pyrophosphate (TPP) that accounts for 80% of the thiamine in tissues [97].

Erythrocyte TPP concentration is influenced by a number of physiological (erythrocyte age and alcohol consumption), pathological [diabetes, liver disorders, polyneuritis, uremic neuropathy, gastrointestinal (GI) disorders, and obesity], and iatrogenic factors (drugs used to treat cancer) [98].

Biochemical assessment of thiamine status

Free thiamine is the main form in the circulation, while TPP predominates in the erythrocytes. The method based on the activation coefficient or ratio (α) of erythrocyte transketolase activity without and with added thiamine developed by Vuilleumier et al. [99] has extensively been used to assess vitamin B_1 nutritional status, and is still a valid "bioassay." An ratio of [enzyme activity (with excess coenzyme)]/[basal enzyme activity (without enzyme)] of 1 or a percentage of zero indicate an adequate thiamine status. The greater the thiamine depletion, the higher the ratios and percentages.

The erythrocyte TPP concentration is recognized as being an adequate indicator of body stores, as its depletion closely parallels those of other principal organs [100]. However, the clinical utility of the EKT functional test is weakened by factors other than the thiamine deficiency per se, including altered enzyme kinetics or levels, compounded with standardization issues and sample storage instability [101,102]. Hence, direct determination of RBC TPP is a more sensitive index of thiamine status. Sensitive, specific, precise, and robust HPLC methods for erythrocyte TPP have been developed over the last 30 years with results correlating to the EKT assay [103]. Urine, easy to collect and noninvasive, can also be considered for assessing recent dietary thiamine intake [98]. Tears have also been used to measure thiamine as well as other water- and fat-soluble vitamins by mass spectrometry [104]. Being a minimally invasive

available biological fluid, it is likely that its use could increase in clinical practice. Total thiamine concentration is expected to be within 60−120 μg/L, where 10% is found in plasma and the remaining 90% as the phosphate ester form in erythrocytes and leukocytes [103].

Riboflavin

Riboflavin (7,8-dimethyl-10-ribityl-isoalloxazine), originally extracted from milk whey by the chemist Wynter-Blyth in 1879, plays a pivotal role in the proper functioning of aerobic cells. Flavin adenine dinucleotide (FAD) and flavin mononucleotide (FMN), its active forms, participate in mitochondrial electron transfer processes and in energy production by acting as cofactors for a variety of enzymes, such as glutathione reductase, L-gulonolactone oxidase, xanthine oxidase, L-amino oxidase, and nicotinamide adenine dehydrogenase [105]. Flavins also contribute to oxidative stress by producing superoxide radicals and yet are responsible for the reduction of hydroperoxydes and are linked to cell death programing [106].

Despite their wide distribution in meat, fish, vegetables, and fruits, endemic deficiency defined by biochemical analysis has been documented in different groups [107−109]. While insufficient riboflavin intake results in primary deficiency, anorexia, intestinal malabsorption syndromes, liver diseases, chronic alcoholism, advanced age, use of diuretics, and repeated hospitalization are some of the contributory risk factors for secondary dietary riboflavin deficiency. Of importance, riboflavin deficiency is rarely isolated. While inflamed oral mucosal tissue and dry scaly skin represent the common dermatological deficiency symptoms, sensorimotor peripheral neuropathy is the main neurological manifestation [110]. However, as riboflavin deficiency is rarely isolated, the above signs and symptoms are likely to reflect multiple deficiencies of the vitamin B complex. Indeed, severe riboflavin deficiency alters the conversion of vitamin B_6 to its coenzyme forms and of tryptophan to niacin. Furthermore, riboflavin deficiency can possibly cause hyperhomocysteinemia through an abated methylenetetrahydrofolate reductase (MTHFR) activity, an FAD-dependent enzyme [111].

Biochemical assessment of riboflavin status

Originally, riboflavin nutritional status was assessed using microbiological assays to indirectly measure the amount of the vitamin in different biological fluids (plasma, whole blood, and urine). However, these methods were perceived to reflect recent vitamin intake not the body stores. Hence, the in vitro stimulation of the erythrocyte glutathione reductase with exogenous FAD, expressed as an erythrocyte glutathione reductase activation coefficient, was developed in the early 1970s and is still being used in population surveys [112,113]. However, care in interpreting results should be applied, as a lack of standardization has been reported, leading to misleading evaluation of riboflavin deficiency in some populations [114]. Care must also be applied before transposing data obtained from a presumably healthy population to subjects with pathologies. Since erythrocyte glutathione reductase activity depends on the cellular FAD concentration, it follows that the direct measurement of FAD in red cells would be a reliable biomarker for a longitudinal assessment of the riboflavin nutritional status, as shown by Hustad et al. [115]. Riboflavin status may also be assessed by measuring urinary excretion of the vitamin in either a random specimen or 24-hour collection, particularly in replete individuals, as it is not a sensitive marker at low vitamin intakes [105,116]. Moreover, there is a call for caution, as there are few data on the effect of kidney malfunction on water-soluble vitamins.

A number of methods for measuring plasma, whole blood and urine riboflavin, FMN, and FAD based on HPLC coupled to fluorometric detection or tandem mass-spectrometry have been published [117−119]. None, however, are based on international reference material as standard, thereby weakening the transferability of results between laboratories and methods. Petteys and Frank [119] using an HPLC−UV method reported a range of riboflavin values in plasma from 6.7 to 50.1 nM.

Niacin

Association of niacin deficiency, referring collectively to nicotinic acid (pyridine-3-carboxylic acid) and its nicotinamide derivative, to pellagra became apparent by the mid-20th century, in populations consuming low-quality protein corn-based diets. The explanation lies in that the preformed nicotinic acid, present in corn, is tightly bound to a protein, preventing its absorption by the small intestine [120]. The classical symptoms of pellagra include sun-sensitive dermatitis, diarrhea, neurological disorders, accompanied by anxiety, insomnia, disorientation, hallucinations, and delirium [121]. Niacin is peculiar among the "vitamin B complex," as it can be obtained with poor efficiency through the conversion of tryptophan, an indispensable amino acid and precursor of serotonin [122]. This process requires multiple factors (pyridoxine, riboflavin, iron, and copper). Hence, conditions in which conversion of tryptophan to nicotinic acid is impaired, such as in isoniazid treatment for tuberculosis that depletes pyridoxine and pyridoxamine stores, or malignant carcinoid syndrome in which tryptophan is preferentially converted to serotonin, or still in Hartnup's syndrome in which tryptophan absorption is impaired, may all lead to niacin

deficiency. Niacin deficiency is also associated with Crohn's disease and chronic alcoholism [123].

Nicotinamide adenine dinucleotide (NAD) and nicotinamide adenine dinucleotide phosphate (NADP) are the major active components of niacin, acting both as cofactors for multiple enzymes involved in redox reactions and energy metabolism and as substrates for poly (ADP-ribosylation) of nuclear proteins. In addition to its action as a vitamin, nicotinic acid (rather than nicotinamide) at pharmacologic doses lowers total cholesterol, LDL-cholesterol, triglyceride, and lipoprotein(a) concentrations, while raising HDL-cholesterol concentrations. These effects, however, have a downside, such as skin flushing, GI distress, hepatotoxicity, gout, and decreased insulin sensitivity [124,125].

Biochemical assessment of niacin status

Early assessment of niacin status involved the determination of the urinary metabolites *N*-methyl-2-pyridone-5-carboxamide and *N*-methylnicotinamide and expressing them as a ratio, the *N*-methyl-2-pyridone metabolite being more sensitive to a low dietary intake than the *N*-methylnicotinamide. In this context, a ratio of <1.0 is indicative of niacin deficiency [126]. It was later demonstrated that the RBC NAD pool decreased rapidly during niacin deficiency, whereas the NADP pool was quite stable. This led Jacobson et al. [127] to suggest NAD/NADP, referred to as the niacin number, as an index of niacin deficiency in humans. Methods for measuring nicotinic acid and its metabolites have evolved with time from the microbiological bioassay using *Leuconostoc mesenteroides*, an obligate user of nicotinic acid [128]. A few years later Huff and Perlzweig [129] developed a sensitive fluorometric assay for *N*′-methylnicotinamide based on its condensation with either acetone or acetophenone, thus producing a fluorescent adduct. HPLC coupled to UV detection [130] or mass spectrometry [131] and electron-capture negative-ion gas chromatography–mass spectrometry [132] have replaced these time-consuming methods with the advantage of being able to provide a metabolic profile of the niacin derivatives. According to the Mayo Clinic, normal ranges for niacin range from 0.50 to 8.91 mg/L [133].

Pantothenic acid

Pantothenic acid, consisting of pantoic acid linked to β-alanine through an amide linkage, is inactive as such. It must be sequentially metabolized to pantothenic acid-4′-phosphate, whose condensation with a cysteine moiety in an ATP-dependent reaction forms 4′-phosphopantothenoylcysteine. Following a decarboxylation and addition of adenosine 5′-monophosphate to 4′-phosphopantetheine, followed by phosphorylation of the ribose 3′-hydroxyl, the active form of the vitamin is produced: coenzyme A (CoA) [134], acting as an acyl carrier for a variety of enzymes involved in lipid synthesis and intermediate metabolism [135].

In humans, pantothenic acid supply depends on the diet and the colonic microflora. Dietary pantothenic acid consists mainly of CoA, from which it has to be hydrolyzed prior to a sodium-dependent multivitamin transporter (SMVT)-mediated absorption in the ileum and the colon [88,136]. Clinically, overt pantothenic acid deficiency has been observed experimentally in individuals fed a low pantothenic acid diet virtually devoid of pantothenic acid. The subjects exhibited varying degrees of irritability, low blood pressure, fatigue; apathy; malaise; sleep disturbances; nausea, vomiting, and abdominal cramps; and numbness, paresthesia, muscle cramps, and staggering gait [137]. Inadequacy of vitamins and minerals, including pantothenic acid, based on food recall diaries, has recently been reported among Moroccan high-school pupils. However, no clinical data were recorded [138].

Biochemical assessment of pantothenic acid status

Reports on the assessment of pantothenic acid nutritional status in humans are scarce and are often related to clinical trials or epidemiological nutritional studies rather than to clinical settings. Pantothenic acid nutritional status is best reflected by whole blood or erythrocyte concentrations, although urine concentration is also highly correlated with dietary intake [87]. It must be borne in mind that pantothenic acid, in cells containing mitochondria, is present mostly as CoA [135]. Therefore an enzymatic digestion is required if total pantothenic acid was to be measured. On the contrary, mature RBC, plasma, and urine pantothenic acid are essentially in the free form. The first record of a microbiological assay for blood and urine pantothenic acid dates back to 1962 [139]. Guilarte [140] later developed a radiometric microbiological assay based on the measurement of $^{14}CO_2$ generated from the metabolism of a ^{14}C-labeled pantothenic acid by yeast *Kloeckeru brevis* in proportion to the amount of vitamin present. This microorganism has an absolute requirement for pantothenic acid for its growth.

Wyse et al. [141] described a radioimmunoassay based on a rabbit polyclonal antibody against pantothenic acid–bovine serum albumin conjugate. The assay demonstrated good specificity, but only moderate correlation with Hatano's microbiological assay, referred to above. It was later used to assess the pantothenic acid nutritional status of a small group of adolescents [142]. More recently,

Takahashi et al. [143] described a rapid, accurate, and precise isocratic reversed-phase ion-pair HPLC method coupled to postcolumn derivatization for the determination of pantothenic acid in human urine. This method was successfully applied for the assessment of the pantothenic acid nutritional status of a group of Japanese pregnant women [144]. Methods using gas chromatography coupled to mass fragmentation of trimethylsilyl derivatives and multiple ion detections have also been published [145,146]. Rychlik [147], on his part, developed a gas chromatography−stable-isotope dilution−mass spectrometry method for quantification of total and free pantothenic acid trimethylsilyl derivatives in food and plasma. These methods being time-consuming and labor-intensive may, however, not be applicable for routine testing in clinical laboratories. There are no reports on the measurement of pantothenic acid in biological samples by HPLC−tandem mass spectrometry, although methods have been developed for its measurement in food and supplements [148,149]. Adaptation of this approach to plasma and urine is an attractive alternative to the other methods, as it would require less intensive sample preparation. Wittwer et al. [150] reported a normal range from 1.57 to 2.66 μmol/L for whole blood pantothenic acid.

Vitamin B_6 (pyridoxine, pyridoxamine, and pyridoxal)

3-Hydroxy-2-methylpyridine is the cyclical core of all B_6 vitamers: pyridoxine, pyridoxamine, pyridoxal, and their phosphorylated counterparts: pyridoxine-5′-phosphate, pyridoxal-5′-phosphate (PLP), pyridoxamine-5′-phosphate, and 4-pyridoxic acid (4-PA). The coenzyme form of vitamin B_6, PLP, is found covalently bound to enzymes through a Schiff base with a specific ε-aminolysine group (close to or within the active site) [151,152]. PLP is the principal active cofactor for amino acid transferases, decarboxylases, and dehydratases. In addition, it is involved in heme biosynthesis through the δ-aminolevulinate synthase, glycogenolysis through the phosphorylase, brain glutamate metabolism, and sphingoid base biosynthesis. Leklem [153] and Chawla and Kvarnberg [154] have extensively reviewed its natural sources, metabolism, absorption mechanisms, and multiple roles in metabolic processes.

As a summary, in view of its central role in numerous biological functions, deficiency in vitamin B_6 usually occurs in combination with the other B vitamins. Restricted intake (particularly infants and the elderly), intestinal malabsorption conditions, such as celiac disease, Crohn's disease, and ulcerative colitis, and chronic alcoholism are the most frequent reasons for its deficiency. Chronic renal insufficiency, hemodialysis, liver disease, sickle cell anemia, rheumatoid arthritis, hyperoxaluria types I and II, and certain types of medications (interference with its bioavailability) are also risk factors for pyridoxine deficiency. Riboflavin deficiency also negatively impacts vitamin B_6 status, since the conversion of pyridoxamine-5′-phosphate and pyridoxine-5′-phosphate to PLP necessitates the FMN-requiring oxidase [155]. PLP deficiency may include cardiovascular, neurologic, GI, dermatologic, and immunological disorders. The cardiovascular risk link can be explained by the fact that PLP is a central cofactor in the one-carbon metabolic pathway and is a coenzyme for cystathionine β-synthase and cystathionine γ-lyase in the transsulfuration pathway [156], thus a deficiency state could impair homocysteine turnover, leading to hyperhomocysteinemia, which is recognized as an independent risk factor for reversible coronary heart disease [157]. Toxicity symptoms related to pyridoxine are rarely reported. They occur upon long-term ingestion of large doses of pyridoxine, and usually consist of motor and peripheral sensory neuropathy, unsteady gait, and photosensitivity expressed as vesicular dermatosis upon sun exposure [153].

Biochemical assessment of vitamin B_6 status

A variety of direct, indirect, and dietary methods have been developed to assess vitamin B_6 nutritional status in humans. Direct indices define those in which one or more of the vitamin B_6 metabolites are measured in plasma, erythrocytes, or urine samples. Indirect methods refer to those requiring measurement of PLP-dependent enzyme activities in biological fluids, usually plasma in the presence and absence of excess PLP. Dietary methods must be interpreted with caution, as they are the least reliable because of the inherent problems in obtaining accurate dietary recalls.

Keeping clinical laboratories in mind, direct methods should be prioritized and circulating PLP concentration is to be considered. Alternatively, 24-hour collection or spot sample urinary 4-PA can also be a useful biomarker. However, these indices are to be interpreted in the proper context; dietary intake of vitamin B_6 and protein, and physiological and pathological variables influence the fasting plasma PLP concentration and urine 4-PA [158−161]. Other indices include the measurement of urinary xanthurenic acid excretion following a tryptophan load test, most useful for monitoring an individual's response to pyridoxine supplementation, and the indirect erythrocyte aspartate aminotransferase activity coefficients (ratio of stimulated/basal activity), and plasma homocysteine [162]. Recently published methods for measuring PLP and 4-PA are based on HPLC coupled to fluorescence detection or mass spectrometry.

Different microbiological and enzymatic methods have been described for the measurement of plasma PLP [163,164]. More recently, HPLC-based methods with fluorescence detection are preferred. However, these methods classically required derivatization of PLP to enhance the fluorescence signal. Cabo et al. [165] recently published an HPLC method without postcolumn derivatization, coupled to a fluorescence detector for the measurement of plasma or serum PLP and 4-PA after precipitation of proteins with trichloroacetic acid. The authors reported a 100% recovery and interassay coefficient of variations of $\leq 6\%$ for both compounds. The development of user-friendly tandem mass spectrometers, coupled to HPLC systems, enables clinical laboratories to use this technology. van Zelst and de Jonge [166] have developed a high-throughput method using this technology. Its advantages are the short chromatography runtime, the high sensitivity and specificity, and the wide concentration spectrum. In this study, reported values ranged from 35 to 110 mM using 120 whole blood samples.

Biotin

Biotin, playing an essential role in the expression of genes involved in intermediate metabolism immune function and cell proliferation [167,168], is required by all classes of living organisms. Biotin (IUPAC name: 5-[(3a*S*,4*S*,6a*R*)-2-oxohexahydro-1*H*-thieno[3,4-*d*]imidazol-4-yl]pentanoic acid) is composed of a tetrahydroimidizalone ring fused with a tetrahydrothiophene ring in a boat conformation with a valeric acid attached to carbon 4 of the tetrahydrothiophene ring. Prior to the Goldberg−Sternbach stereospecific synthesis pathway [169], the chemical synthesis of biotin yields eight enantiomers. Of these, only one designated **d**-(+)-biotin is enzymatically active. Humans, like other mammalians, cannot synthesize biotin and thus must obtain it from exogenous sources. As for a number of other water-soluble vitamins, two sources of biotin are available: a dietary source, biotin being widely distributed in foods; and the normal microflora present in the colon. Dietary protein-bound biotin, converted by GI proteases, peptidases, and biotinidase to free biotin [90,170], is absorbed in the small and large intestines through a saturable and Na^+-dependent carrier-mediated process, exclusively expressed at the apical membrane of enterocytes, that is shared with pantothenic acid and lipoic acid [88]. The human intestinal biotin uptake process is adaptively upregulated in biotin deficiency via transcriptionally mediated mechanisms that involve Kruppel-like factor 4 sites [170].

Biotin is a coenzyme for at least six biotin-dependent carboxylases, two decarboxylases, and one transcarboxylase that include pyruvate carboxylase, methylcrotonyl-CoA carboxylase, propionyl-CoA carboxylase, and acetyl-CoA carboxylase [171]. Biotin is conjugated to these apoenzymes through the action of the enzyme holocarboxylase synthetase, driven by the transfer of the adenylate portion of ATP, with the formation of biotinyl-5′-adenylate and pyrophosphate that is immediately hydrolyzed by pyrophosphatase [168]. Dietary deficiency is rare; however, it can occur in individuals chronically consuming raw egg whites that contain avidin, a high-affinity biotin-binding protein, patients on total parenteral nutrition given formulas without biotin, and in inborn errors of metabolism causing multiple carboxylase deficiencies or biotinidase deficiency. Biotin deficiency includes organic aciduria associated with neurologic abnormalities comprising developmental delay in infants, hypotonia, impaired consciousness, seizures, ataxia, depression, and cutaneous lesions (rash and alopecia) [168,172,173]. However, the clinical presentation and age of onset of symptoms are variable.

Biochemical assessment of vitamin biotin status

Traditional methods of quantifying biotin in biological fluids include microbiological assays and bioassay, and have already been reviewed [173,174]. Other methods depend on competitive binding assays of biotin to either avidin or streptavidin after a prior purification of biotin and its metabolites by HPLC [175]. Circulating biotin (serum or plasma) has also been proposed as an index of the biotin status. However, it proved not to be a predictable biomarker, because biotin deficiency was not always reflected by low circulating biotin [174]. More recently, quantification of urinary 3-hydroxyisovaleric acid (3-HIA), an intermediate that accumulates and is excreted upon decreased methylcrotonyl-CoA carboxylase activity during biotin deficiency, has furthered the advancement of the assessment of biotin status in the early process of biotin deficiency [176,177].

Different methods have been developed to measure urinary 3-HIA and related compounds, including gas chromatography−mass spectrometry analysis after di-trimethylsilyl derivatization [178], reversed-phase HPLC and UV detection at 230 nm of the 2-nitrophenylhydrazine derivative [179], and ultrahigh-performance liquid chromatography−tandem mass spectrometry (UPLC−MS/MS) [177]. Urinary biotin average values of 0.74 and 0.52 nM biotin/μM creatinine were reported by Eng et al. [176] using avidin-binding assay for biotin sufficient and biotin insufficient, respectively.

Folic acid

Folic acid is a collective term designating pteroylglutamic acids and their oligoglutamic acid conjugates: *N*-[p-[[(2-amino-4-hydroxypteridin-6-yl)methyl]amino]benzoyl]-L

(+)-(n)-glutamic acid. Commonly, the number of glutamic acid residues varies between 3 and 7. It is an important cofactor, which, together with vitamin B_{12}, plays a key role in the utilization of methyl groups, essential in cell division and multiple metabolic reactions [180]. Furthermore, folic acid is involved in DNA biosynthesis, and hence participates in the regulation of gene expression and transcription, chromatin structure, and genomic stability [181]. In the one-carbon cycle, MTHFR catalyzes the reduction of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate, the main circulating form of folate. Transference of the 5-methyltetrahydrofolate methyl group to homocysteine, mediated by methionine synthase, yields methionine to which an adenosyl residue is added to form S-adenosylmethionine, a universal methyl donor [180].

In addition to folate, the metabolism of homocysteine and methionine requires vitamin B_{12}, vitamin B_6, and vitamin B_2 as cofactors for the numerous enzyme reactions. Adequate supply of these vitamins and fully operational enzymes assure normal homocysteine levels. Elevated circulating homocysteine concentration is an independent risk factor for cardiovascular disease and colorectal cancer [182]. Delvin et al. [183] have shown that plasma homocysteine concentration correlates negatively with folate and vitamin B_{12} concentrations and that MTHFR genotype plays a significant role in determining homocysteine concentrations, particularly in nutritionally stressed children older than 10 years. The mechanisms of folate absorption, metabolism, and excretion have already been reviewed [87]. In brief, dietary folate polyglutamates are hydrolyzed in the gut by a γ-glutamyl hydrolase to a monoglutamate form before being actively transported across the proximal small intestine by a saturable pH-dependent process. A nonsaturable passive diffusion takes over when pharmacological doses of folate monoglutamate are consumed. The liver, where it is metabolized to polyglutamate derivatives and retained or released into the blood or bile, takes much of the folate present in the portal circulation. Approximately, two-thirds of the folate in plasma is protein-bound, with albumin accounting for 50% of protein binding. Cellular transport of folate is mediated by membrane carriers and folate-binding proteins. Before being stored in tissue or used as a coenzyme in single-carbon transfer reactions, folate monoglutamate must be converted to the polyglutamate form and reduced enzymatically. When released from tissues into circulation, it is reconverted to the monoglutamate form [180].

Megaloblastic anemia, defined by RBC macrocytosis and abnormal nuclear maturation, is a cardinal feature of folate deficiency and is indistinguishable from that caused by vitamin B_{12} deficiency, since the two vitamins are substrates or cofactors for the methionine synthase reaction and other pathways, leading to DNA precursor synthesis. Discrimination between the two is crucial for an effective treatment. Although folate deficiency symptoms are usually due to a dietary insufficiency, chronic excessive alcohol consumption, malabsorption syndromes, or iatrogenic causes must be considered in the clinical workup. Also, an increased requirement may be due to genetic variation including women who require increased folate in early pregnancy to reduce the risk of neural tube defects [184]. The fortification of cereal and grain products, required in the United States since 1998, resulted in a documented rise in the serum and RBC folate levels, and successfully decreased the incidence of neural tube defects in neonates and of megaloblastic anemia. Widespread food fortification, and the between-laboratory and intermethod analytical variations question the clinical utility of folate measurements to detect folate deficiency. Gilfix, examining the literature, revealed that, in clinical situations, in general, $<1\%$ of the subjects will have folate deficiency regardless of the assay method used and of potentially predisposing factors [185].

Biochemical assessment of vitamin folate status

Assessment of folate falls into two categories: short- and long-term nutritional statuses. Whereas plasma folate (essentially monoglutamate) relates the recent intake, RBC polyglutamate, however, reflects the average content of the circulating red cell population, and thus could be considered the "glycated hemoglobin" equivalent to diabetes monitoring. Caution should be exercised in interpreting results in the presence of abnormal hemoglobin with shorter half-life, such as in sickle cell anemia. As for plasma folate, adequately spaced serial plasma measurements are needed to allow the distinction between acute and chronic folic acid deficiencies. Whereas a single plasma folate measurement is not reliable in a nutrition intervention, particularly short term, it, however, provides a fitting assessment in population surveys. Plasma folate consists almost entirely of 5-methyltetrahydrofolate (5-methyl-THF), and RBC of long-chain polyglutamates of 5-methyl-THF. Microbiological assays, classified as "functional assays" using *Lactobacillus casei*, were used first to assess folate status and served to establish the present-day reference values and cutoffs, defining the deficiency status. Methods based on HPLC coupled to electrochemical detection have also been developed [186] and have been used in clinical settings. However, due to their relative simplicity of use, assays, using high-affinity folate-binding proteins to compete with endogenous and labeled folic acid (or methyltetrahydrofolate), and using either radioactive, enzyme-linked, or chemiluminescent detection systems, have gradually supplanted microbiological assays in clinical laboratories [187].

The question raised by Yetley and Johnson [188] as to whether differences in folate concentrations observed from one study to another are due to population changes or are methodology artifacts has been answered by Colapinto et al. [189] who have shown that the difference between the Canadian and American RBC folate concentrations was assay method-dependent. An interlaboratory "round robin" comparison study published in 1996 that included 20 participants revealed, for serum and whole blood, overall within-method coefficients of variation of 27.6% and 35.7%, respectively, and two- to ninefold differences in concentrations between methods, with the greatest variation occurring at critical low folate concentrations [190]. These studies clearly show the need for harmonization, and preferably standardization of the assay methods against an international reference method traceable to the SRM 1955 developed by NIST [191]. As a number of such methods, based on HPLC or UPLC−MS/MS, have been published in the last 10 years [192−194], a second "round robin" study would be timely. Also, worthy of note, the preanalytical conservation protocols are crucial, since delayed processing and freezing of the plasma samples result in clinically unacceptable folate losses [195,196]. Fazili et al. [193] reported a range of total folic acid (including five folate vitamers) ranging from 10 to 100 nM in serum, where average folic acid levels were 2 nM.

Vitamin B_{12}

As already mentioned, vitamin B_{12} and folic acid have common metabolic pathways that define the size of the methyl donor pool utilized in multiple metabolic pathways. Vitamin B_{12} in its natural form is present only in foods of animal origin, explaining why its deficiency is more common among strict vegetarians and populations consuming little animal food. Detailing the chemistry pathways and the physiology of vitamin B_{12} is beyond the scope of this chapter; more detailed reviews can be consulted [197−199].

The generic term cobalamin (vitamin B_{12}) refers to structures, named corrin, consisting of one central cobalt (Co) atom coordinated with four equatorial nitrogen atoms donated by pyrrole residues. The fifth Co coordination site (α), below the planar structure, is occupied by a 5′, 6′-dimethyl benzimidazole residue linked to an α-ribosyl-3-phosphate. The sixth Co coordination site (β), above the corrin ring, is occupied either by a methyl [methylcobalamin (MeCbl)], 5′-deoxyadenosyl [5′-deoxyadenosylcobalamin (AdoCbl)], or a cyanide group (CNCbl) that has no biological activity. MeCbl is the cofactor of several methyltransferases, such as methionine synthase, which catalyzes the conversion of homocysteine to methionine; AdoCbl is the cofactor of enzymes such as L-methylmalonyl-CoA mutase that catalyzes the reversible isomerization of L-methylmalonate into the succinate.

Cobalamin synthesis is a complex process restricted to certain strains of bacteria, such as *Lactobacillus rossiae* and *Lactobacillus reuteri*, which, present in the human gut microbiota, are sources of the vitamin [200,201]. Red meat is another source of cobalamin, accounting for 20%−40% of the dietary intake [202]. Paradoxically, epidemiological studies recommend a limited consumption of red meat to reduce risk of obesity, cancer, and other chronic diseases, as they contain a fair content of saturated fat. To circumvent the health risk, the consumption of lean red meat has been proposed as a healthy alternative because of its low saturated fat content [203]. After intestinal uptake, cobalamin circulates in the bloodstream bound to two carriers: 20% to one of three transcobalamins (TCs I, II, and III) named holotranscobalamin (holoTC), and the remainder to the metabolically inert haptocorrin [199]. Holo II is the form that has biological function allowing the delivery of the vitamin to the target cells. Having a short serum half-life, it is accepted as being the earliest indicator of negative vitamin B_{12} balance, its concentration falling before serum total vitamin B_{12}.

Following ingestion, the release of vitamin B_{12} from food is dependent upon stomach acidity, explaining in part the higher prevalence of vitamin B_{12} deficiency in the elderly population. Other conditions such as intestinal malabsorption, postgastric/intestinal surgery, or the loss of the distal ileum can also contribute to vitamin B_{12} deficiency. Since vitamin B_{12} absorption is dependent on binding to the protein intrinsic factor (IF) in the small intestine, its absence leads to malabsorption and pernicious anemia [199]. Megaloblastic anemia is the prevalent effect of vitamin B_{12} (and folate) deficiency, followed by neurological effects that include tingling and numbness, abnormalities in walking, and cognitive changes ranging from simple memory loss to severe dementia [204]. Decreased energy, shortness of breath, loss of appetite, constipation, glossitis, and papillary atrophy of the tongue are also associated with vitamin B_{12} deficiency.

Biochemical assessment of vitamin B_{12} status

No single test adequately allows the diagnosis of vitamin B_{12} deficiency. The Schilling test was the first to be described to measure vitamin B_{12} absorption. However, the administration of radioactively labeled cobalamin that is needed to follow the vitamin in the circulation and in the urine has limited its use in recent years. Present-day biochemical tests used to assess vitamin B_{12} status include serum total vitamin B_{12}, TC II, and methylmalonic acid (MMA). Despite its low sensitivity and specificity, serum vitamin B_{12} is the first-line test when vitamin

B_{12} deficiency is suspected. As for other vitamins, serum vitamin B_{12} was initially measured by microbiological methods using, for example, *Lactobacillus leichmannii*, for which cobalamin is an obligate nutrient. These methods, being functional assays, had the advantage of measuring the biologically active cobalamin fraction, and hence were more accurate. They, however, suffered from interlaboratory variability, interference by antibiotics, and extended turnaround time inherent to the assay principle [205]. The microbiological assays were gradually replaced by nonspecific R-protein-binding assays. However, the values obtained with these were markedly higher than those obtained with the microbiological functional assay, as they measured cobalamin analogs that are not biologically active [206]. This problem has been resolved by the introduction of IF-based radioassays that compared well with the microbiological assays [207]. The development of immunochemistry analyzers led clinical laboratories to opt for high-throughput automated competitive binding chemiluminescence assays, using purified IF as reagent, that measure total cobalamin after its release from the endogenous binding proteins. Some caution should be observed, however, in interpreting results, as reports have shown that some of these assays are vulnerable to the presence of interfering anti-IF antibodies, particularly in patients with pernicious anemia, thereby providing spuriously elevated serum cobalamin concentrations in otherwise vitamin B_{12}-deficient patients [208,209].

The long-standing lack of standardization against internationally accepted reference materials for serum vitamin B_{12} measurements has led to intermethodology variability that complicates interstudy comparisons. The World Health Organization Expert Committee on Biological Standardization established, in 1992, a first international standard for serum vitamin B_{12} (coded 81/563) that was assayed using a variety of methods [210]. Also, the NIST is currently developing an isotope dilution−liquid chromatography−tandem mass spectrometric (ID−LC−MS/MS) method for vitamin B_{12} (and vitamin B_6), to validate the standard reference material SRM 3951, consisting of three levels of vitamin B_{12} spanning the range usually found in patient samples. This initiative will improve the accuracy of marketed routine cobalamin tests and reduce differences in laboratory proficiency [211]. Aside from the analytical dilemma, cutoffs are customarily derived from reference intervals obtained from different groups, representative or not of a specific population, and as these are method- and population-dependent, it follows that characterizing cobalamin status remains challenging. This is of public health interest, as the prevalence of "cobalamin deficiency" will vary with different cutoff points. Bailey et al. [212] proved the point by reporting that, in NHANES surveys, the prevalence of cobalamin deficiency ranged from 2.9% at a cutoff of <148 pmol/L to 25.7% at <258 pmol/L.

HoloTC II, the biologically active form of vitamin B_{12}, has been proposed as a replacement of total cobalamin. A high-throughput automated two-step sandwich microparticle enzyme immunoassay allows its measurement in clinical settings [213]. However, its clinical utility has been challenged in a thorough review reporting a marginally better receiver operating curve analysis when MMA, a functional indicator of vitamin B_{12} status, is taken as reference [214]. The problem is compounded by the variable MMA cutoff points accepted as indicative of vitamin B_{12} deficiency [215]. The same argument holds for total cobalamin, and, in both cases, the variable cutoffs led to a wide variation in the prevalence of vitamin B_{12} deficiency. Recently, Bailey et al. [216] modeled the relationship between the serum vitamin B_{12} and MMA concentrations. Fig. 46.4 shows that the cutoff for vitamin B_{12} deficiency is at ≈ 120 pmol/L.

Plasma homocysteine has also been proposed as a functional indicator for vitamin B_{12} and folate status, although its sensitivity is questioned [214]. As it can be appreciated, any biomarker has its own limitation, and none qualifies as a single diagnostic standard. Hence, a combination of biomarkers is needed to delineate vitamin B_{12} nutritional status, and an evaluation of algorithms is needed to circumscribe this issue. According to the NIH MedlinePlus [217], the expected reference intervals for vitamin B_{12} range between 200 and 900 ng/L, values that Bailey et al. [212] argued could lead to significant misclassifications.

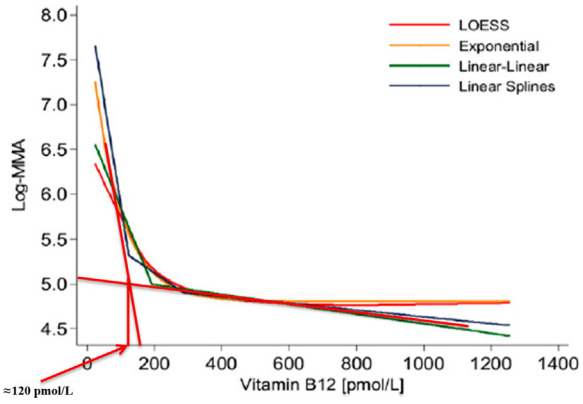


FIGURE 46.4 Relationship between serum methylmalonic acid and vitamin B_{12} concentrations. Loess, exponential, linear−linear, and linear−spline models of the relation between serum vitamin B_{12} and log-methylmalonic acid concentrations among US adults aged >19 years, 1999−2004. Tangent lines were hand free drawn. The intercept and projection lines to the *x*-axis give a threshold estimate for vitamin B_{12} deficiency of 120 pmol/L. *Adapted with permission from R.L. Bailey, R.A. Durazo-Arvizu, R. Carmel, R. Green, C.M. Pfeiffer, C.T. Sempos, et al., Modeling a methylmalonic acid-derived change point for serum vitamin B-12 for adults in NHANES, Am. J. Clin. Nutr. 98 (2) (2013) 460−467.*

Choline

Although described in the late 19th century, choline was officially recognized as an essential micronutrient for humans in 1998 when the Food and Nutrition Board of the United States of America Institute of Medicine of the National Academy of Sciences issued guidelines on its daily intake [218]. Choline (trimethyl-beta-hydroxyethylammonium) is widely distributed in plants and animals. Choline is essential for the synthesis of methionine and carnitine and ACh, and is a building block of phosphatidylcholine and sphingomyelin, important elements of cell membranes. As a methyl donor choline is the third member of the triumvirate that includes folic acid and vitamin B_{12} involved in epigenetic gene regulation through DNA and histone methylation [219]. In terms of bioavailability, choline can be obtained endogenously through de novo synthesis, although this pathway is not sufficient to sustain an adequate supply [218]. Water-soluble choline, phosphocholine, and glycerophosphocholine are absorbed via the portal circulation, and the widely food-distributed lipid-soluble phosphatidylcholine and sphingomyelin are absorbed into lymph as chylomicrons. Although most individuals have adequate choline intake, pregnant and lactating women, infants, cirrhotic and anorexic patients, and those on parenteral nutrition are at higher risk of deficiency. Choline deficiency has been associated with nonalcoholic fatty liver disease and muscle dysfunction [220]. Excessive choline intake from food has been associated with body odor, sweating, salivation, hypotension, and hepatotoxicity in humans [218].

Biochemical assessment of choline status

Although methods for measuring free and phospholipid-bound choline are available, requests for choline measurement addressed to clinical laboratories are infrequent. The first radiolabelling labor-intensive method for measuring free choline developed by Wang and Haubrich [221] from the Squibb Institute was rapidly replaced by nonisotopic methods. Takayama et al. [222] published an enzymatic method to measure choline-containing phospholipids that was used once to assess serum-free choline and phospholipid-bound choline concentrations in pregnancy and in newborns [223]. Other methods described were based on micropyrolysis gas chromatography [224], and more recently on stable-isotope dilution–ultraperformance liquid chromatography–mass spectrometry (UPLC–MS/MS), which allows a metabolic profiling by simultaneous quantifying ACh, betaine, choline, and dimethylglycine [225]. Although sensitive, precise, and rapid, the use of these methods has so far been limited in clinical investigation. However, the development of increasingly user-friendly mass spectrometers will allow application of these methods for clinical evaluation. This is of importance, since high frequency of inadequate choline intake has recently been reported in pregnant women enrolled in the Programming of Intergenerational Stress Mechanisms study [226]. Kirsch et al. [225] reported plasma choline values of 6.99–12.35 μM and 1.14–4.7 mM choline/mol creatinine in urine in 54 adults of >50 years.

Vitamin C

The biosynthetic pathway of vitamin C (ascorbic acid) differs in plants and fungi from mammals. In the first group, it is initiated from mannose and in the second from glucose. Both pathways, however, need the oxidation of L-gulonolactone by L-gulonolactone oxidase, which is not functional in primates, including the human species [227,228]. Vitamin C is involved in biological systems as varied as neurotransmission, peptide hormone synthesis and secretion, and extracellular matrix integrity. It is an obligate cofactor for dopamine beta-hydroxylase that catalyzes norepinephrine synthesis, and of peptidylglycine amidating monooxygenase that catalyzes the conversion of proenkephalin to metenkephalin in chromaffin vesicles. It is also an essential actor in collagen fibril maturation, through the hydroxylation of lysine and proline residues required for collagen chain and fibril crosslinking in a variety of connective tissues. It also acts as a cofactor for nucleic acid and histone demethylation, proteoglycan deglycanation, and acts as both a prooxidant and an antioxidant [229].

Scurvy is the hallmark of severe and prolonged vitamin C deficiency, of which the, first “Modern” description, is attributed to Lind in his treatise published in 1753 [230], leading to supplementation through lime consumption. Despite its early recognition, scurvy remained a problem for explorers into the beginning of the 20th century. While frank scurvy, characterized by subcutaneous and intramuscular hemorrhages, leg edema, neuropathy, and cerebral hemorrhage, is uncommon in industrialized countries, vitamin C insufficiency may be more prevalent than generally assumed despite the widespread availability of vitamins. A UK Low Income Diet and Nutrition Survey based on a representative sample of the low-income population estimated that 25% of men and 16% of women had plasma vitamin C concentrations indicative of deficiency [231]. Persons with excessive alcohol consumption, those consuming diets containing little or no fruits and vegetables, who present with skin rashes and oral lesions, and complain of muscle weakness and pain are likely to be vitamin C-deficient [230]. The symptoms associated with scurvy are explained by the central role that ascorbic acid plays in the numerous biological systems mentioned earlier.

Early reports have associated long-term vitamin C megadose treatment with nephrolithiasis (oxaluria), diarrhea, nausea, dental erosion, and hemolysis in the presence of glucose-6-phosphate dehydrogenase deficiency, an X-linked enzymopathy [232,233]. However, other reports and pharmacokinetic studies challenge these observations [234]. Finally, as pointed out by Lucock et al. [235], the possible synergetic interaction between dietary ascorbic acid and folate should not be overlooked when large doses of vitamins are administered. Cigarette smoking, oral contraceptive, and chronic aspirin use, as well as age, gender, pregnancy, and infection, all influence circulating ascorbic acid concentration and tissue stores [236].

Biochemical assessment of vitamin C status

Vitamin C nutritional assessment is most frequently based on the measure of serum and urine (preferably 24-hour collection) ascorbic acid concentrations. Fasting conditions are essential, because foods are commonly supplemented with the ascorbate epimer erythorbic acid, which is not antiscorbutic. Serum and urine ascorbic acid concentrations reflect the most recent intake of vitamin C, which for leukocytes correlates well with tissue stores. However, white blood cells, as well as erythrocytes, whole blood, saliva, and oral cavity ascorbic acid contents, are used essentially in research protocols, these assays being more complex and time-consuming [236]. Stable isotope-labeled ascorbic acid (^{13}C) dilution has been shown to be useful for the measurement of vitamin C absorption and body pool size in human subjects [237] but is of little use in clinical settings.

Vitamin C can be measured by a variety of methods, including spectrophotometry, colorimetry, capillary electrophoresis, and microdialysis. Independently of the method, preanalytical precautions are, however, paramount for reliable results. Ascorbic acid measurement in biological samples is highly dependent on the type of sampling method (EDTA versus heparin vacuum tubes), rapidity of processing (immediate centrifugation and plasma acidification), storage conditions, and length of time (avoiding light, $-70^{\circ}C$ without delay, and <80 days) [238,239]. After 1980, the use of HPLC-based assays followed by UV, fluorescence, or electrochemical detection became the standard, improving analytical specificity and sensitivity [240–242]. Fluorescence detection offers higher sensitivity but requires postcolumn derivatization with a fluorophore (O-phenylenediamine). Electrochemical detection is thought to provide the best measurement, because it differentiates between ascorbic acid and Derythorbic acid when nonfasting samples are inadvertently used [243]. Jansen and Ruskovska [73] reported plasma vitamin C reference intervals ranging between 4.1 and 20 mg/L.

References

[1] R.T.C.V.S. Lagua, Nutrition and Diet Therapy Reference Dictionary, fifth ed., Blackwell Publishing Ltd, Oxford, UK, 2004, p. 403.

[2] E. Reboul, P. Borel, Proteins involved in uptake, intracellular transport and basolateral secretion of fat-soluble vitamins and carotenoids by mammalian enterocytes, Prog. Lipid Res. 50 (4) (2011) 388–402.

[3] Y. Yamanashi, T. Takada, R. Kurauchi, Y. Tanaka, T. Komine, H. Suzuki, Transporters for the intestinal absorption of cholesterol, vitamin E, and vitamin K, J. Atheroscler. Thromb 24 (4) (2017) 347–359.

[4] E. Reboul, Vitamin E bioavailability: mechanisms of intestinal absorption in the spotlight, Antioxidants (Basel) 6 (4) (2017).

[5] N. Peretti, A. Sassolas, C.C. Roy, C. Deslandres, M. Charcosset, J. Castagnetti, et al., Guidelines for the diagnosis and management of chylomicron retention disease based on a review of the literature and the experience of two centers, Orphanet J. Rare Dis. 5 (2010) 24.

[6] IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN), Nomenclature of retinoids. Recommendations 1982, Eur. J. Biochem. 123 (3) (1982) 473–475.

[7] H. Chen, W.N. Howald, M.R. Juchau, Biosynthesis of all-trans-retinoic acid from all-trans-retinol: catalysis of all-trans-retinol oxidation by human P-450 cytochromes, Drug Metab. Dispos. 28 (3) (2000) 315–322.

[8] R. Blomhoff, H.K. Blomhoff, Overview of retinoid metabolism and function, J. Neurobiol. 66 (7) (2006) 606–630.

[9] P.S. Park, Constitutively active rhodopsin and retinal disease, Adv. Pharmacol. 70 (2014) 1–36.

[10] J.C. Sherwin, M.H. Reacher, W.H. Dean, J. Ngondi, Epidemiology of vitamin A deficiency and xerophthalmia in at-risk populations, Trans. R. Soc. Trop. Med. Hyg. 106 (4) (2012) 205–214.

[11] C.C. Brown, R.J. Noelle, Seeing through the dark: New insights into the immune regulatory functions of vitamin A, Eur. J. Immunol. 45 (5) (2015) 1287–1295.

[12] K.L. Penniston, S.A. Tanumihardjo, The acute and chronic toxic effects of vitamin A, Am. J. Clin. Nutr. 83 (2) (2006) 191–201.

[13] E. Chroni, A. Monastirli, D. Tsambaos, Neuromuscular adverse effects associated with systemic retinoid dermatotherapy: monitoring and treatment algorithm for clinicians, Drug Saf. 33 (1) (2010) 25–34.

[14] V.S. Ramanathan, G. Hensley, S. French, V. Eysselein, D. Chung, S. Reicher, et al., Hypervitaminosis A inducing intra-hepatic cholestasis–a rare case report, Exp. Mol. Pathol. 88 (2) (2010) 324–325.

[15] K.H. Safi, A.G. Filbrun, S.Z. Nasr, Hypervitaminosis A causing hypercalcemia in cystic fibrosis. Case report and focused review, Ann. Am. Thorac. Soc. 11 (8) (2014) 1244–1247.

[16] E.H. Harrison, Mechanisms involved in the intestinal absorption of dietary vitamin A and provitamin A carotenoids, Biochim. Biophys. Acta 1821 (1) (2012) 70–77.

[17] S.A. Tanumihardjo, Vitamin A: biomarkers of nutrition for development, Am. J. Clin. Nutr. 94 (2) (2011) 658S–665SS.

[18] W.J. Driskell, J.S. Hewett, M.M. Bashor, Evaluation of a direct fluorometric method for determination of serum retinol, Clin. Chem. 32 (5) (1986) 867–869.

[19] A. Semeraro, I. Altieri, M. Patriarca, A. Menditto, Evaluation of uncertainty of measurement from method validation data: an application to the simultaneous determination of retinol and alpha-tocopherol in human serum by HPLC, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 877 (11−12) (2009) 1209−1215.

[20] D. Thibeault, H. Su, E. MacNamara, H.M. Schipper, Isocratic rapid liquid chromatographic method for simultaneous determination of carotenoids, retinol, and tocopherols in human serum, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 877 (11−12) (2009) 1077−1083.

[21] J.B. Thomas, D.L. Duewer, I.O. Mugenya, K.W. Phinney, L.C. Sander, K.E. Sharpless, et al., Preparation and value assignment of standard reference material 968e fat-soluble vitamins, carotenoids, and cholesterol in human serum, Anal. Bioanal. Chem. 402 (2) (2012) 749−762.

[22] S.L. Arnold, J.K. Amory, T.J. Walsh, N. Isoherranen, A sensitive and specific method for measurement of multiple retinoids in human serum with UHPLC-MS/MS, J. Lipid Res. 53 (3) (2012) 587−598.

[23] R. Rizzoli, Nutritional aspects of bone health, Best. Pract. Res. Clin. Endocrinol. Metab. 28 (6) (2014) 795−808.

[24] E. Delvin, J.C. Souberbielle, J.P. Viard, B. Salle, Role of vitamin D in acquired immune and autoimmune diseases, Crit. Rev. Clin. Lab. Sci. 51 (4) (2014) 232−247.

[25] S. Wang, Epidemiology of vitamin D in health and disease, Nutr. Res. Rev. 22 (2) (2009) 188−203.

[26] A. Prentice, Nutritional rickets around the world, J. Steroid Biochem. Mol. Biol. 136 (2013) 201−206.

[27] E. Delvin, N. Alos, F. Rauch, V. Marcil, S. Morel, M. Boisvert, et al., Vitamin D nutritional status and bone turnover markers in childhood acute lymphoblastic leukemia survivors: a PETALE study, Clin. Nutr. (2018).

[28] J.M. Pettifor, A. Prentice, The role of vitamin D in paediatric bone health, Best. Pract. Res. Clin. Endocrinol. Metab. 25 (4) (2011) 573−584.

[29] P. Lips, N.M. van Schoor, The effect of vitamin D on bone and osteoporosis, Best. Pract. Res. Clin. Endocrinol. Metab. 25 (4) (2011) 585−591.

[30] M.G. Vogiatzi, E. Jacobson-Dickman, M.D. DeBoer, Vitamin D supplementation and risk of toxicity in pediatrics: a review of current literature, J. Clin. Endocrinol. Metab. 99 (4) (2014) 1132−1141.

[31] Institute of Medicine Committee to Review Dietary Reference Intakes for Vitamin D, Calcium, The National Academies Collection: Reports Funded by National Institutes of Health, in: A.C. Ross, C.L. Taylor, A.L. Yaktine, H.B. Del Valle (Eds.), Dietary Reference Intakes for Calcium and Vitamin D, National Academies Press, National Academy of Sciences, Washington, DC, 2011.

[32] M.J. Glade, A 21st century evaluation of the safety of oral vitamin D, Nutrition 28 (4) (2012) 344−356.

[33] G. Conti, V. Chirico, A. Lacquaniti, L. Silipigni, C. Fede, A. Vitale, et al., Vitamin D intoxication in two brothers: be careful with dietary supplements, J. Pediatr. Endocrinol. Metab. 27 (7−8) (2014) 763−767.

[34] C. Chambellan-Tison, B. Horen, G. Plat-Wilson, P. Moulin, I. Claudet, Severe hypercalcemia due to vitamin D intoxication, Arch. Pediatr. 14 (11) (2007) 1328−1332.

[35] S.H. Allen, J.H. Shah, Calcinosis and metastatic calcification due to vitamin D intoxication. A case report and review, Horm. Res. 37 (1−2) (1992) 68−77.

[36] A.B. Liborio, J.C. Nasserala, A.S. Gondim, E.F. Daher, The case: renal failure in a bodybuilder athlete. Diagnosis: nephrocalcinosis secondary to exogenous vitamin D intoxication, Kidney Int. 85 (5) (2014) 1247−1248.

[37] H.L. Henry, Regulation of vitamin D metabolism, Best. Pract. Res. Clin. Endocrinol. Metab. 25 (4) (2011) 531−541.

[38] G. Jones, Expanding role for vitamin D in chronic kidney disease: importance of blood 25-OH-D levels and extra-renal 1alpha-hydroxylase in the classical and nonclassical actions of 1alpha,25-dihydroxyvitamin D(3), Semin. Dial 20 (4) (2007) 316−324.

[39] P. Lips, Vitamin D deficiency and secondary hyperparathyroidism in the elderly: consequences for bone loss and fractures and therapeutic implications, Endocr. Rev. 22 (4) (2001) 477−501.

[40] M.J. Shearer, P. Newman, Recent trends in the metabolism and cell biology of vitamin K with special reference to vitamin K cycling and MK-4 biosynthesis, J. Lipid Res. 55 (3) (2014) 345−362.

[41] J.F. Aloia, S.A. Talwar, S. Pollack, M. Feuerman, J.K. Yeh, Optimal vitamin D status and serum parathyroid hormone concentrations in African American women, Am. J. Clin. Nutr. 84 (3) (2006) 602−609.

[42] L. Steingrimsdottir, O. Gunnarsson, O.S. Indridason, L. Franzson, G. Sigurdsson, Relationship between serum parathyroid hormone levels, vitamin D sufficiency, and calcium intake, JAMA 294 (18) (2005) 2336−2341.

[43] P. Patel, M.Z. Mughal, P. Patel, B. Yagnik, N. Kajale, R. Mandlik, et al., Dietary calcium intake influences the relationship between serum 25-hydroxyvitamin D3 (25OHD) concentration and parathyroid hormone (PTH) concentration, Arch. Dis. Child. 101 (4) (2016) 316−319.

[44] C.T. Sempos, H.W. Vesper, K.W. Phinney, L.M. Thienpont, P.M. Coates, Vitamin D status as an international issue: national surveys and the problem of standardization, Scand. J. Clin. Lab. Invest. Suppl. 243 (2012) 32−40.

[45] D. Stockl, P.M. Sluss, L.M. Thienpont, Specifications for trueness and precision of a reference measurement system for serum/plasma 25-hydroxyvitamin D analysis, Clin. Chim. Acta 408 (1−2) (2009) 8−13.

[46] N. Binkley, C.T. Sempos, Standardizing vitamin D assays: the way forward, J. Bone Miner. Res. 29 (8) (2014) 1709−1714.

[47] K. Sarafin, R. Durazo-Arvizu, L. Tian, K.W. Phinney, S. Tai, J.E. Camara, et al., Standardizing 25-hydroxyvitamin D values from the Canadian Health Measures Survey, Am. J. Clin. Nutr. (2015).

[48] C. Le Goff, E. Cavalier, J.C. Souberbielle, A. González-Antuña, E. Delvin, Measurement of circulating 25-hydroxyvitamin D: a historical review, Pract. Lab. Med. 2 (2015) 1−14.

[49] D. Blanton, Z. Han, L. Bierschenk, M.V. Linga-Reddy, H. Wang, M. Clare-Salzler, et al., Reduced serum vitamin D-binding protein levels are associated with type 1 diabetes, Diabetes 60 (10) (2011) 2566−2570.

[50] S. Malik, L. Fu, D.J. Juras, M. Karmali, B.Y. Wong, A. Gozdzik, et al., Common variants of the vitamin D binding protein gene and adverse health outcomes, Crit. Rev. Clin. Lab. Sci. 50 (1) (2013) 1−22.

[51] J.R. Delanghe, R. Speeckaert, M.M. Speeckaert, Behind the scenes of vitamin D binding protein: more than vitamin D binding, Best. Pract. Res. Clin. Endocrinol. Metab. 29 (2015).
[52] G. Jones, M. Kaufmann, Vitamin D metabolite profiling using liquid chromatography-tandem mass spectrometry (LC-MS/MS), J. Steroid Biochem. Mol. Biol. 164 (2016) 110–114.
[53] J.M.W. van den Ouweland, Analysis of vitamin D metabolites by liquid chromatography-tandem mass spectrometry, Trends Analyt. Chem. 84 (2016) 117–130.
[54] M.E. Jensen, F.M. Ducharme, Y. Theoret, A.S. Belanger, E. Delvin, Data in support for the measurement of serum 25-hydroxyvitamin D (25OHD) by tandem mass spectrometry, Data Brief. 8 (2016) 925–929.
[55] E. Delvin, M. Boisvert, M.A. Lecours, Y. Théorêt, M. Kaufmann, G. Jones, et al., Data for the measurement of serum vitamin D metabolites in childhood acute lymphoblastic leukemia survivors, Data Brief. 18 (2018) 1427–1432.
[56] M.F. Holick, N.C. Binkley, H.A. Bischoff-Ferrari, C.M. Gordon, D. A. Hanley, R.P. Heaney, et al., Evaluation, treatment, and prevention of vitamin D deficiency: an Endocrine Society clinical practice guideline, J. Clin. Endocrinol. Metab. 96 (7) (2011) 1911–1930.
[57] Mustacich D.J., Bruno R.S., Traber M.G. Vitamin E. Vitam. Horm. 2007;76:1-21.
[58] M.G. Traber, Determinants of plasma vitamin E concentrations, Free. Radic. Biol. Med. 16 (2) (1994) 229–239.
[59] M.G. Traber, J. Atkinson, Vitamin E, antioxidant and nothing more, Free Radic. Biol. Med. 43 (1) (2007) 4–15.
[60] M.G. Traber, R.J. Sokol, G.W. Burton, C. Kiyose, T. Ueda, O. Igarashi, et al., Impaired ability of patients with familial isolated vitamin E deficiency to incorporate alpha-tocopherol into lipoproteins secreted by the liver, J. Clin. Invest 85 (2) (1990) 397–407.
[61] A. Hosomi, M. Arita, Y. Sato, K.U. Ingold, A.M. Papas, J.E. Huffaker, et al., Affinity for alpha-tocopherol transfer protein as a determinant of the biological activities of vitamin E analogs, FEBS Lett. 409 (1) (1997) 105–108.
[62] I. Di Donato, S. Bianchi, A. Federico, Ataxia with vitamin E deficiency: update of molecular diagnosis, Neurol. Sci. 31 (4) (2010) 511–515.
[63] G. El Euch-Fayache, Y. Bouhlal, R. Amouri, M. Feki, F. Hentati, Molecular, clinical and peripheral neuropathy study of Tunisian patients with ataxia with vitamin E deficiency, Brain 137 (Pt 2) (2014) 402–410.
[64] K. Saito, S. Matsumoto, T. Yokoyama, M. Okaniwa, S. Kamoshita, Pathology of chronic vitamin E deficiency in fatal familial intrahepatic cholestasis (Byler disease), Virchows Arch. A. Pathol. Anat. Histol 396 (3) (1982) 319–330.
[65] V. Kalra, J. Grover, G.K. Ahuja, S. Rathi, D.S. Khurana, Vitamin E deficiency and associated neurological deficits in children with protein-energy malnutrition, J. Trop. Pediatr 44 (5) (1998) 291–295.
[66] W.H. Oldewage-Theron, F.O. Samuel, R.D. Djoulde, Serum concentration and dietary intake of vitamins A and E in low-income South African elderly, Clin. Nutr. 29 (1) (2010) 119–123.
[67] D.K. Dror, L.H. Allen, Vitamin E deficiency in developing countries, Food Nutr. Bull. 32 (2) (2011) 124–143.
[68] W.P. James, M. Nelson, A. Ralph, S. Leather, Socioeconomic determinants of health. The contribution of nutrition to inequalities in health, BMJ 314 (7093) (1997) 1545–1549.
[69] M. Rutkowski, K. Grzegorczyk, Adverse effects of antioxidative vitamins, Int. J. Occup. Med. Environ. Health 25 (2) (2012) 105–121.
[70] M.G. Traber, Vitamin E and K interactions--a 50-year-old problem, Nutr. Rev. 66 (11) (2008) 624–629.
[71] C. Yuan, M. Burgyan, D.R. Bunch, E. Reineks, R. Jackson, R. Steinle, et al., Fast, simple, and sensitive high-performance liquid chromatography method for measuring vitamins A and E in human blood plasma, J. Sep. Sci. 37 (17) (2014) 2293–2299.
[72] E. Hinchliffe, J. Rudge, P. Reed, A novel high-throughput method for supported liquid extraction of retinol and alpha-tocopherol from human serum and simultaneous quantitation by liquid chromatography tandem mass spectrometry, Ann. Clin. Biochem. (2015).
[73] E. Jansen, T. Ruskovska, Serum biomarkers of (anti)oxidant status for epidemiological studies, Int. J. Mol. Sci. 16 (11) (2015) 27378–27390.
[74] A. Goncalves, S. Roi, M. Nowicki, A. Dhaussy, A. Huertas, M.J. Amiot, et al., Fat-soluble vitamin intestinal absorption: absorption sites in the intestine and interactions for absorption, Food Chem. 172 (2015) 155–160.
[75] P. Krzyzanowska, S. Drzymala-Czyz, A. Pogorzelski, M. Dus-Zuchowska, W. Skorupa, L. Bober, et al., Vitamin K status in cystic fibrosis patients with liver cirrhosis, Dig. Liver Dis. 49 (6) (2017) 672–675.
[76] D.J. Card, R. Gorska, J. Cutler, D.J. Harrington, Vitamin K metabolism: current knowledge and future research, Mol. Nutr. Food Res. 58 (8) (2014) 1590–1600.
[77] M.S. Hamidi, O. Gajic-Veljanoski, A.M. Cheung, Vitamin K and bone health, J. Clin. Densitom. 16 (4) (2013) 409–413.
[78] B.A. Willems, C. Vermeer, C.P. Reutelingsperger, L.J. Schurgers, The realm of vitamin K dependent proteins: shifting from coagulation toward calcification, Mol. Nutr. Food Res. 58 (8) (2014) 1620–1635.
[79] G. Ferland, Vitamin K, an emerging nutrient in brain function, Biofactors 38 (2) (2012) 151–157.
[80] O.W. Charles, B.C. Dilworth, E.J. Day, Chick bioassay of vitamin K compounds using dicumarol and pivalyl as anticoagulants, Poult. Sci. 47 (3) (1968) 754–760.
[81] E.J. Favaloro, G. Lippi, The new oral anticoagulants and the future of haemostasis laboratory testing, Biochem. Med. 22 (3) (2012) 329–341.
[82] K.W. Davidson, J.A. Sadowski, Determination of vitamin K compounds in plasma or serum by high-performance liquid chromatography using postcolumn chemical reduction and fluorimetric detection, Methods Enzymol 282 (1997) 408–421.
[83] I. Meineke, C. Willnow, J. Stingl, J. Brockmoller, Routine micromethod for the determination of vitamin K1 in human plasma, Ther. Drug Monit. 35 (6) (2013) 859–862.
[84] V. Ducros, M. Pollicand, F. Laporte, A. Favier, Quantitative determination of plasma vitamin K1 by high-performance liquid chromatography coupled to isotope dilution tandem mass spectrometry, Anal. Biochem. 401 (1) (2010) 7–14.
[85] A. Gentili, A. Cafolla, T. Gasperi, S. Bellante, F. Caretti, R. Curini, et al., Rapid, high performance method for the determination of vitamin K(1), menaquinone-4 and vitamin K(1) 2,3-epoxide in human serum and plasma using liquid chromatography-hybrid quadrupole linear ion trap mass spectrometry, J. Chromatogr. A 1338 (2014) 102–110.

[86] Mayo Medical Laboratories VK. Available from: <http://www.mayomedicallaboratories.com/test-catalog/Clinical + and + Interpretive/62167>.

[87] Standing Committee on the Scientific Evaluation of Dietary Reference Intakes and its Panel on Folate OV, and Choline and Subcommittee on Upper Reference Levels of Nutrients, Food and Nutrition Board, Institute of Medicine, Pantothenic acid. Dietary Reference Intakes for Thiamine, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic acid, Biotin, and Choline, The National Academies Press, Washington, DC, 1998, pp. 357−373.

[88] H.M. Said, Intestinal absorption of water-soluble vitamins in health and disease, Biochem. J. 437 (3) (2011) 357−372.

[89] H.M. Said, Z.M. Mohammed, Intestinal absorption of water-soluble vitamins: an update, Curr. Opin. Gastroenterol 22 (2) (2006) 140−146.

[90] H.M. Said, Recent advances in transport of water-soluble vitamins in organs of the digestive system: a focus on the colon and the pancreas, Am. J. Physiol. Gastrointest. Liver Physiol 305 (9) (2013) G601−G610.

[91] P.R. Kiela, F.K. Ghishan, Physiology of Intestinal Absorption and Secretion, Best. Pract. Res. Clin. Gastroenterol 30 (2) (2016) 145−159.

[92] C.M. Clase, V. Ki, R.M. Holden, Water-soluble vitamins in people with low glomerular filtration rate or on dialysis: a review, Semin. Dial 26 (5) (2013) 546−567.

[93] E. Descombes, O. Boulat, F. Perriard, G. Fellay, Water-soluble vitamin levels in patients undergoing high-flux hemodialysis and receiving long-term oral postdialysis vitamin supplementation, Artif. Organs 24 (10) (2000) 773−778.

[94] J. Elliott, D. Mishler, R. Agarwal, Hyporesponsiveness to erythropoietin: causes and management, Adv. Chronic Kidney Dis. 16 (2) (2009) 94−100.

[95] K.J. Carpenter, The discovery of thiamine, Ann. Nutr. Metab. 61 (3) (2012) 219−223.

[96] A.S. Hazell, R.F. Butterworth, Update of cell damage mechanisms in thiaminee deficiency: focus on oxidative stress, excitotoxicity and inflammation, Alcohol Alcohol 44 (2) (2009) 141−147.

[97] B. Shaanan, D.M. Chipman, Reaction mechanisms of thiamine diphosphate enzymes: new insights into the role of a conserved glutamate residue, FEBS J. 276 (9) (2009) 2447−2453.

[98] G. Rs, Assessment of the status of thiamine, riboflavin and niacin, in: G. Rs (Ed.), Principles of Nutritional Assessment, second ed., Oxford University Press, New York, NY, 2005, pp. 545−574.

[99] J.P. Vuilleumier, H.E. Keller, R. Rettenmaier, F. Hunziker, Clinical chemical methods for the routine assessment of the vitamin status in human populations. Part II: the water-soluble vitamins B_1, B_2 and B_6, Int. J. Vitam. Nutr. Res. 53 (4) (1983) 359−370.

[100] L.G. Warnock, C.R. Prudhomme, C. Wagner, The determination of thiamine pyrophosphate in blood and other tissues, and its correlation with erythrocyte transketolase activity, J. Nutr. 108 (3) (1978) 421−427.

[101] A.L. Bailey, P.M. Finglas, A.J. Wright, S. Southon, Thiamine intake, erythrocyte transketolase (EC 2.2.1.1) activity and total erythrocyte thiamine in adolescents, Br. J. Nutr. 72 (1) (1994) 111−125.

[102] J.A. Puxty, A.E. Haskew, J.G. Ratcliffe, J. McMurray, Changes in erythrocyte transketolase activity and the thiaminee pyrophosphate effect during storage of blood, Ann. Clin. Biochem. 22 (Pt 4) (1985) 423−427.

[103] D. Talwar, H. Davidson, J. Cooney, D. St JO'Reilly, Vitamin B (1) status assessed by direct measurement of thiamine pyrophosphate in erythrocytes or whole blood by HPLC: comparison with erythrocyte transketolase activation assay, Clin. Chem. 46 (5) (2000) 704−710.

[104] M. Khaksari, L.R. Mazzoleni, C. Ruan, R.T. Kennedy, A.R. Minerick, Determination of water-soluble and fat-soluble vitamins in tears and blood serum of infants and parents by liquid chromatography/mass spectrometry, Exp. Eye Res. 155 (2017) 54−63.

[105] H.J. Powers, Riboflavin (vitamin B-2) and health, Am. J. Clin. Nutr. 77 (6) (2003) 1352−1360.

[106] A.A. Heikal, Intracellular coenzymes as natural biomarkers for metabolic activities and mitochondrial anomalies, Biomark. Med. 4 (2) (2010) 241−263.

[107] F. Rohner, M.B. Zimmermann, R. Wegmueller, A.B. Tschannen, R.F. Hurrell, Mild riboflavin deficiency is highly prevalent in school-age children but does not increase risk for anaemia in Cote d'Ivoire, Br. J. Nutr. 97 (5) (2007) 970−976.

[108] M.E. Keith, N.A. Walsh, P.B. Darling, A. McClelland, E. Madraa, A.B. Chandia, et al., B-vitamin deficiency in hospitalized patients with heart failure, J. Am. Diet. Assoc. 109 (8) (2009) 1406−1410.

[109] E.K. Nichols, L.E. Talley, N. Birungi, S.A. Hanninen, S. Thirugnanam, H. Leong-Poi, et al., Suspected outbreak of riboflavin deficiency among populations reliant on food assistance: a case study of drought-stricken Karamoja, Uganda, 2009−2010, PLoS One 8 (5) (2013) e62976.

[110] D.C.M.D. Heimburger, M.E. Shils, Clinical manifestations of nutrient deficiencies and toxicities: a resume, in: S. Me (Ed.), Modern Nutrition in Health and Disease, 10th ed., Williams and Wilkins, Philadelphia, 2006, pp. 595−612.

[111] M.L. Martinez-Frias, The biochemical structure and function of methylenetetrahydrofolate reductase provide the rationale to interpret the epidemiological results on the risk for infants with Down syndrome, Am. J. Med. Genet. Part. A 146a (11) (2008) 1477−1482.

[112] G.E. Nichoalds, Assessment of status riboflavin nutriture by assay of erythrocyte glutathione reductase activity, Clin. Chem. 20 (5) (1974) 624−628.

[113] K.C. Whitfield, C.D. Karakochuk, Y. Liu, A. McCann, A. Talukder, H. Kroeun, et al., Poor thiamine and riboflavin status is common among women of childbearing age in rural and urban Cambodia, J. Nutr. 145 (3) (2015) 628−633.

[114] M.H. Hill, A. Bradley, S. Mushtaq, E.A. Williams, H.J. Powers, Effects of methodological variation on assessment of riboflavin status using the erythrocyte glutathione reductase activation coefficient assay, Br. J. Nutr. 102 (2) (2009) 273−278.

[115] S. Hustad, M.C. McKinley, H. McNulty, J. Schneede, J.J. Strain, J.M. Scott, et al., Riboflavin, flavin mononucleotide, and flavin adenine dinucleotide in human plasma and erythrocytes at

baseline and after low-dose riboflavin supplementation, Clin. Chem. 48 (9) (2002) 1571–1577.

[116] M.D. Riboflavin, in: M.E. Shils, Sm, A.C. Ross, B. Caballero, R. J. Cousins (Eds.), Modern Nutrition in Health and Disease, 10th ed., Lippincott Williams & Wilkins, Philadelphia, 2006, pp. 434–441.

[117] O. Midttun, S. Hustad, E. Solheim, J. Schneede, P.M. Ueland, Multianalyte quantification of vitamin B_6 and B_2 species in the nanomolar range in human plasma by liquid chromatography-tandem mass spectrometry, Clin. Chem. 51 (7) (2005) 1206–1216.

[118] S.S. Patil, A.K. Srivastava, Development and validation of a liquid chromatography method for the simultaneous determination of eight water-soluble vitamins in multivitamin formulations and human urine, J. AOAC Int. 96 (6) (2013) 1273–1280.

[119] B.J. Petteys, E.L. Frank, Rapid determination of vitamin B(2) (riboflavin) in plasma by HPLC, Clin. Chim. Acta 412 (1–2) (2011) 38–43.

[120] J. Laguna, K.J. Carpenter, Raw versus processed corn in niacin-deficient diets, J. Nutr. 45 (1) (1951) 21–28.

[121] D.J. Lanska, Chapter 30: historical aspects of the major neurological vitamin deficiency disorders: the water-soluble B vitamins, Handb. Clin. Neurol. 95 (2010) 445–476.

[122] N. Le Floc'h, W. Otten, E. Merlot, Tryptophan metabolism, from nutrition to potential therapeutic applications, Amino Acids 41 (5) (2011) 1195–1205.

[123] J.B. Kirkland, R.J. Niacin, in: R.B.S.J. Rucker, D.B. McCormick, L.J. Machlin (Eds.), Handbook of Vitamins, third ed., Marcel Dekker Inc, New York, NY, 2001, pp. 213–254.

[124] J.C. Creider, R.A. Hegele, T.R. Joy, Niacin: another look at an underutilized lipid-lowering medication, Nat. Rev. Endocrinol. 8 (9) (2012) 517–528.

[125] W.L. Song, G.A. FitzGerald, Niacin, an old drug with a new twist, J. Lipid Res. 54 (10) (2013) 2586–2594.

[126] R.A. Jacob, M.E. Swendseid, R.W. McKee, C.S. Fu, R.A. Clemens, Biochemical markers for assessment of niacin status in young men: urinary and blood levels of niacin metabolites, J. Nutr. 119 (4) (1989) 591–598.

[127] E.L. Jacobson, M.K. Jacobson, N.A.D. Tissue, As a biochemical measure of niacin status in humans, Methods Enzymol 280 (1997) 221–230.

[128] S. Gaines, G.L. Stahly, The growth requirements of leuconostoc mesenteroides and preliminary studies on its use as an assay agent for several members of the vitamin B complex, J. Bacteriol. 46 (5) (1943) 441–449.

[129] J.W. Huff, W.A. Perlzweig, The fluorescent condensation product of N1-methylnicotinamide and acetone; a sensitive method for the determination of N1-methylnicotinamide in urine, J. Biol. Chem. 167 (1) (1947) 157–167.

[130] P.F. Chatzimichalakis, V.F. Samanidou, R. Verpoorte, I.N. Papadoyannis, Development of a validated HPLC method for the determination of B-complex vitamins in pharmaceuticals and biological fluids after solid phase extraction, J. Sep. Sci. 27 (14) (2004) 1181–1188.

[131] R. Mullangi, N.R. Srinivas, Niacin and its metabolites: role of LC-MS/MS bioanalytical methods and update on clinical pharmacology. An overview, Biomed. Chromatogr 25 (1–2) (2011) 218–237.

[132] G.A. Smythe, O. Braga, B.J. Brew, R.S. Grant, G.J. Guillemin, S.J. Kerr, et al., Concurrent quantification of quinolinic, picolinic, and nicotinic acids using electron-capture negative-ion gas chromatography-mass spectrometry, Anal. Biochem. 301 (1) (2002) 21–26.

[133] Mayo Medical Laboratories N. Available from: <http://www.mayomedicallaboratories.com/test-catalog/Clinical + and + Interpretive/91379>.

[134] G.M. Brown, The metabolism of pantothenic acid, J. Biol. Chem. 234 (2) (1959) 370–378.

[135] N. Plesofsky, Pantothenic acid, in: R.B. Sj Rucker, D.B. McCormick, L.J. Machlin (Eds.), Handbook of Vitamins, third ed., Marcel Dekker Inc., New York, NY, 2001, pp. 317–337.

[136] H.M. Said, A. Ortiz, E. McCloud, D. Dyer, M.P. Moyer, S. Rubin, Biotin uptake by human colonic epithelial NCM460 cells: a carrier-mediated process shared with pantothenic acid, Am. J. Physiol. 275 (5 Pt 1) (1998) C1365–C1371.

[137] P.C. Fry, H.M. Fox, H.G. Tao, Metabolic response to a pantothenic acid deficient diet in humans, J. Nutr. Sci. Vitaminol 22 (4) (1976) 339–346.

[138] K. Anzid, A. Baali, P. Vimard, S. Levy-Desroches, M. Cherkaoui, P.M. Lopez, Inadequacy of vitamins and minerals among high-school pupils in Ouarzazate, Morocco, Public Health Nutr. 17 (8) (2014) 1786–1795.

[139] M. Hatano, Microbiological assay of pantothenic acid in blood and urine, J. Vitaminol 8 (1962) 134–142.

[140] T.R. Guilarte, A radiometric microbiological assay for pantothenic acid in biological fluids, Anal. Biochem. 178 (1) (1989) 63–66.

[141] B.W. Wyse, C. Wittwer, R.G. Hansen, Radioimmunoassay for pantothenic acid in blood and other tissues, Clin. Chem. 25 (1) (1979) 108–110.

[142] B.R. Eissenstat, B.W. Wyse, R.G. Hansen, Pantothenic acid status of adolescents, Am. J. Clin. Nutr. 44 (6) (1986) 931–937.

[143] K. Takahashi, T. Fukuwatari, K. Shibata, Fluorometric determination of pantothenic acid in human urine by isocratic reversed-phase ion-pair high-performance liquid chromatography with post-column derivatization, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 877 (22) (2009) 2168–2172.

[144] K. Shibata, T. Fukuwatari, S. Sasaki, M. Sano, K. Suzuki, C. Hiratsuka, et al., Urinary excretion levels of water-soluble vitamins in pregnant and lactating women in Japan, J. Nutr. Sci. Vitaminol 59 (3) (2013) 178–186.

[145] K. Banno, M. Matsuoka, S. Horimoto, J. Kato, Simultaneous determination of pantothenic acid and hopantenic acid in biological samples and natural products by gas chromatography-mass fragmentography, J. Chromatogr. 525 (2) (1990) 255–264.

[146] K. Banno, Measurement of pantothenic acid and hopantenic acid by gas chromatography-mass spectroscopy, Methods Enzymol 279 (1997) 213–219.

[147] M. Rychlik, Quantification of free and bound pantothenic acid in foods and blood plasma by a stable isotope dilution assay, J. Agric. Food Chem. 48 (4) (2000) 1175–1181.

[148] P. Andrieux, P. Fontannaz, T. Kilinc, E.C. Gimenez, Pantothenic acid (vitamin B_5) in fortified foods: comparison of a novel ultra-performance liquid chromatography-tandem mass spectrometry method and a microbiological assay (AOAC Official Method 992.07), J. AOAC Int. 95 (1) (2012) 143–148.

[149] F. Gosetti, B. Bolfi, M. Manfredi, G. Calabrese, E. Marengo, Determination of eight polyphenols and pantothenic acid in extra-virgin olive oil samples by a simple, fast, high-throughput and sensitive ultra high performance liquid chromatography with tandem mass spectrometry method, J. Sep. Sci. (2015).
[150] C.T. Wittwer, C. Schweitzer, J. Pearson, W.O. Song, C.T. Windham, B.W. Wyse, et al., Enzymes for liberation of pantothenic acid in blood: use of plasma pantetheinase, Am. J. Clin. Nutr. 50 (5) (1989) 1072−1078.
[151] D.M. Smith, N.R. Thomas, D. Gani, A comparison of pyridoxal 5'-phosphate dependent decarboxylase and transaminase enzymes at a molecular level, Experientia 47 (11−12) (1991) 1104−1118.
[152] R.C. Hughes, W.T. Jenkins, E.H. Fischer, The site of binding of pyridoxal-5'-phosphate to heart glutamic-aspartic transaminase, Proc. Natl. Acad. Sci. U. S. A. 48 (1962) 1615−1618.
[153] J.E. Leklem, Vitamin B_6, in: R.B. Rucker, Sj, D.B. McCormick, L.J. Machlin (Eds.), Handbooks of Vitamins, third ed., Marcel Dekker Inc, New York, NY, 2001, pp. 339−426.
[154] J. Chawla, D. Kvarnberg, Hydrosoluble vitamins, in: first ed, J.F. J. Biller (Ed.), Neurological Aspects of Systemic Disease, Part II. Handbook of Clinical Neurology, 120, Elsevier B.V., 2014, pp. 891−914.
[155] M.R. Lowik, H. van den Berg, C. Kistemaker, H.A. Brants, J.H. Brussaard, Interrelationships between riboflavin and vitamin B_6 among elderly people (Dutch Nutrition Surveillance System), Int. J. Vitam. Nutr. Res. 64 (3) (1994) 198−203.
[156] J. Selhub, Homocysteine metabolism, Annu. Rev. Nutr. 19 (1999) 217−246.
[157] K. Robinson, E.L. Mayer, D.P. Miller, R. Green, F. van Lente, A. Gupta, et al., Hyperhomocysteinemia and low pyridoxal phosphate. Common and independent reversible risk factors for coronary artery disease, Circulation 92 (10) (1995) 2825−2830.
[158] A. Waldmann, B. Dorr, J.W. Koschizke, C. Leitzmann, A. Hahn, Dietary intake of vitamin B_6 and concentration of vitamin B_6 in blood samples of German vegans, Public Health Nutr. 9 (6) (2006) 779−784.
[159] K. Shibata, A. Tachiki, K. Mukaeda, T. Fukuwatari, S. Sasaki, Y. Jinno, Changes in plasma pyridoxal 5'-phosphate concentration during pregnancy stages in Japanese women, J. Nutr. Sci. Vitaminol 59 (4) (2013) 343−346.
[160] L. Clement, M. Boylan, V. Miller, J. Driskell, D. Giraud, H. Subih, Vitamin B-6 vitamer levels in plasma and related symptoms in hemodialysis subjects taking low- and high-dose renal multivitamin supplements, Int. J. Vitam. Nutr. Res. 82 (2) (2012) 130−136.
[161] C.M. Hansen, J.E. Leklem, L.T. Miller, Changes in vitamin B-6 status indicators of women fed a constant protein diet with varying levels of vitamin B-6, Am. J. Clin. Nutr. 66 (6) (1997) 1379−1387.
[162] G. Rs, Assessment of the vitamin B_6 status, in: G. Rs (Ed.), Principle in Nutritional Assessement, second ed., Oxford University Press, Inc., New York, NY, 2005, pp. 575−594.
[163] T.R. Guilarte, P.A. McIntyre, Radiometric-microbiologic assay fo vitamin B-6: analysis of plasma samples, J. Nutr. 111 (11) (1981) 1861−1868.
[164] Q. Han, M. Xu, L. Tang, X. Tan, X. Tan, Y. Tan, et al., Homogeneous, nonradioactive, enzymatic assay for plasma pyridoxal 5-phosphate, Clin. Chem. 48 (9) (2002) 1560−1564.
[165] R. Cabo, K. Kozik, M. Milanowski, S. Hernes, A. Slettan, M. Haugen, et al., A simple high-performance liquid chromatography (HPLC) method for the measurement of pyridoxal-5-phosphate and 4-pyridoxic acid in human plasma, Clin. Chim. Acta 433 (2014) 150−156.
[166] B.D. van Zelst, R. de Jonge, A stable isotope dilution LC-ESI-MS/MS method for the quantification of pyridoxal-5'-phosphate in whole blood, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 903 (2012) 134−141.
[167] N. Kothapalli, G. Camporeale, A. Kueh, Y.C. Chew, A.M. Oommen, J.B. Griffin, et al., Biological functions of biotinylated histones, J. Nutr. Biochem. 16 (7) (2005) 446−448.
[168] R.J. McMahon, Biotin in metabolism and molecular biology, Annu. Rev. Nutr. 22 (2002) 221−239.
[169] M.W.S.L. Goldberg, Inventor Enantioselective Synthesis of (+)-Biotin, USA, 1949.
[170] H.M. Said, Cell and molecular aspects of human intestinal biotin absorption, J. Nutr. 139 (1) (2009) 158−162.
[171] L. Tong, Structure and function of biotin-dependent carboxylases, Cell Mol. Life Sci 70 (5) (2013) 863−891.
[172] H.M. Said, Biotin: biochemical, physiological and clinical aspects, Subcell. Biochem. 56 (2012) 1−19.
[173] Standing Committee on the Scientific Evaluation of Dietary Reference Intakes and Its Panel on Folate OV, and Choline and Subcommittee on Upper Reference Levels of Nutrients, Food and Nutrition Board, Institute of Medicine, Biotin, in: Sciences NAo (Ed.), Dietary Reference Intakes for Thiamine, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Patothenic Acid, Biotin and Choline, National Academies Press, Washington, DC, 1998, pp. 374−389.
[174] D.M. Mock, Biotin status: which are valid indicators and how do we know? J. Nutr. 129 (2S Suppl) (1999) 498s−503s.
[175] J. Zempleni, D.M. Mock, Advanced analysis of biotin metabolites in body fluids allows a more accurate measurement of biotin bioavailability and metabolism in humans, J. Nutr. 129 (2S Suppl) (1999) 494s−497ss.
[176] W.K. Eng, D. Giraud, V.L. Schlegel, D. Wang, B.H. Lee, J. Zempleni, Identification and assessment of markers of biotin status in healthy adults, Br. J. Nutr. 110 (2) (2013) 321−329.
[177] T.D. Horvath, N.I. Matthews, S.L. Stratton, D.M. Mock, G. Boysen, Measurement of 3-hydroxyisovaleric acid in urine from marginally biotin-deficient humans by UPLC-MS/MS, Anal. Bioanal. Chem. 401 (9) (2011) 2805−2810.
[178] D.M. Mock, H. Jackson, G.L. Lankford, N.I. Mock, S.T. Weintraub, Quantification of urinary 3-hydroxyisovaleric acid using deuterated 3-hydroxyisovaleric acid as internal standard, Biomed. Environ. Mass Spectrom 18 (9) (1989) 652−656.
[179] T. Watanabe, K. Oguchi, S. Ebara, T. Fukui, Measurement of 3-hydroxyisovaleric acid in urine of biotin-deficient infants and mice by HPLC, J. Nutr. 135 (3) (2005) 615−618.
[180] P.J. Stover, Physiology of folate and vitamin B_{12} in health and disease, Nutr. Rev. 62 (6 Pt 2) (2004) S3−S12. discussion S3.
[181] K.S. Crider, T.P. Yang, R.J. Berry, L.B. Bailey, Folate and DNA methylation: a review of molecular mechanisms and the evidence for folate's role, Adv. Nutr. 3 (1) (2012) 21−38.
[182] J.W. Zinck, M. de Groh, A.J. MacFarlane, Genetic modifiers of folate, vitamin B-12, and homocysteine status in a cross-

sectional study of the Canadian population, Am. J. Clin. Nutr. 101 (6) (2015) 1295–1304.
[183] E.E. Delvin, R. Rozen, A. Merouani, J. Genest Jr., M. Lambert, Influence of methylenetetrahydrofolate reductase genotype, age, vitamin B-12, and folate status on plasma homocysteine in children, Am. J. Clin. Nutr. 72 (6) (2000) 1469–1473.
[184] E.H. Reynolds, The neurology of folic acid deficiency, Handb. Clin. Neurol. 120 (2014) 927–943.
[185] B.M. Gilfix, Utility of measuring serum or red blood cell folate in the era of folate fortification of flour, Clin. Biochem. 47 (7–8) (2014) 533–538.
[186] T. Opladen, V.T. Ramaekers, G. Heimann, N. Blau, Analysis of 5-methyltetrahydrofolate in serum of healthy children, Mol. Genet. Metab. 87 (1) (2006) 61–65.
[187] R. Green, Indicators for assessing folate and vitamin B-12 status and for monitoring the efficacy of intervention strategies, Am. J. Clin. Nutr. 94 (2) (2011) 666s–672ss.
[188] E.A. Yetley, C.L. Johnson, Folate and vitamin B-12 biomarkers in NHANES: history of their measurement and use, Am. J. Clin. Nutr. 94 (1) (2011) 322s–331ss.
[189] C.K. Colapinto, M.S. Tremblay, S. Aufreiter, T. Bushnik, C.M. Pfeiffer, D.L. O'Connor, The direction of the difference between Canadian and American erythrocyte folate concentrations is dependent on the assay method employed: a comparison of the Canadian Health Measures Survey and National Health and Nutrition Examination Survey, Br. J. Nutr. 112 (11) (2014) 1873–1881.
[190] E.W. Gunter, B.A. Bowman, S.P. Caudill, D.B. Twite, M.J. Adams, E.J. Sampson, Results of an international round robin for serum and whole-blood folate, Clin. Chem. 42 (10) (1996) 1689–1694.
[191] B.C. Nelson, C.M. Pfeiffer, M. Zhang, D.L. Duewer, K.E. Sharpless, K.A. Lippa, Commutability of NIST SRM 1955 homocysteine and folate in frozen human serum with selected total homocysteine immunoassays and enzymatic assays, Clin. Chim. Acta 395 (1–2) (2008) 99–105.
[192] B.C. Nelson, M.B. Satterfield, L.T. Sniegoski, M.J. Welch, Simultaneous quantification of homocysteine and folate in human serum or plasma using liquid chromatography/tandem mass spectrometry, Anal. Chem. 77 (11) (2005) 3586–3593.
[193] Z. Fazili, R.D. Whitehead Jr., N. Paladugula, C.M. Pfeiffer, A high-throughput LC-MS/MS method suitable for population biomonitoring measures five serum folate vitamers and one oxidation product, Anal. Bioanal. Chem. 405 (13) (2013) 4549–4560.
[194] F. Kiekens, J. Van Daele, D. Blancquaert, D. Van Der Straeten, W.E. Lambert, C.P. Stove, A validated ultra-high-performance liquid chromatography-tandem mass spectrometry method for the selective analysis of free and total folate in plasma and red blood cells, J. Chromatogr. A 1398 (2015) 20–28.
[195] N.P. Kubasik, M. Graham, H.E. Sine, Storage and stability of folate and vitamin B-12 in plasma and blood samples, Clin. Chim. Acta 95 (1) (1979) 147–149.
[196] Z. Fazili, M.R. Sternberg, C.M. Pfeiffer, Assessing the influence of 5,10-methylenetetrahydrofolate reductase polymorphism on folate stability during long-term frozen storage, thawing, and repeated freeze/thawing of whole blood, Clin. Chim. Acta 413 (11–12) (2012) 966–972.
[197] L. Randaccio, S. Geremia, N. Demitri, J. Wuerges, Vitamin B_{12}: unique metalorganic compounds and the most complex vitamins, Molecules 15 (5) (2010) 3228–3259.
[198] B. Krautler, Biochemistry of B_{12}-cofactors in human metabolism, Subcell. Biochem. 56 (2012) 323–346.
[199] C. Gherasim, M. Lofgren, R. Banerjee, Navigating the B(12) road: assimilation, delivery, and disorders of cobalamin, J. Biol. Chem. 288 (19) (2013) 13186–13193.
[200] M. De Angelis, F. Bottacini, B. Fosso, P. Kelleher, M. Calasso, R. Di Cagno, et al., Lactobacillus rossiae, a vitamin B_{12} producer, represents a metabolically versatile species within the genus *Lactobacillus*, PLoS One 9 (9) (2014) e107232.
[201] J.G. LeBlanc, C. Milani, G.S. de Giori, F. Sesma, D. van Sinderen, M. Ventura, Bacteria as vitamin suppliers to their host: a gut microbiota perspective, Curr. Opin. Biotechnol 24 (2) (2013) 160–168.
[202] S. Sharma, T. Sheehy, L.N. Kolonel, Contribution of meat to vitamin B(1)(2), iron and zinc intakes in five ethnic groups in the USA: implications for developing food-based dietary guidelines, J. Hum. Nutr. Diet. 26 (2) (2013) 156–168.
[203] A.J. McAfee, E.M. McSorley, G.J. Cuskelly, B.W. Moss, J.M. Wallace, M.P. Bonham, et al., Red meat consumption: an overview of the risks and benefits, Meat Sci. 84 (1) (2010) 1–13.
[204] W. Herrmann, R. Obeid, Cobalamin deficiency, Subcell. Biochem. 56 (2012) 301–322.
[205] H. Baker, O. Frank, I. Pasher, H. Sobotka, Vitamin B_{12} in human blood and serum. I. Comparison of microbiologic assays using normal subjects, Clin. Chem. 6 (1960) 578–581.
[206] J.F. Kolhouse, H. Kondo, N.C. Allen, E. Podell, R.H. Allen, Cobalamin analogues are present in human plasma and can mask cobalamin deficiency because current radioisotope dilution assays are not specific for true cobalamin, N. Engl. J. Med. 299 (15) (1978) 785–792.
[207] N.P. Kubasik, M. Ricotta, H.E. Sine, Commercially-supplied binders for plasma cobalamin (vitamin B_{12}), analysis--"purified" intrinsic factor, "cobinamide"-blocked R-protein binder, and non-purified intrinsic factor-R-protein binder--compared to microbiological assay, Clin. Chem. 26 (5) (1980) 598–600.
[208] R. Carmel, Y.P. Agrawal, Failures of cobalamin assays in pernicious anemia, N. Engl. J. Med. 367 (4) (2012) 385–386.
[209] M.J. Oberley, D.T. Yang, Laboratory testing for cobalamin deficiency in megaloblastic anemia, Am. J. Hematol. 88 (6) (2013) 522–526.
[210] S.J. Thorpe, The development and role of international biological reference materials in the diagnosis of anaemia, Biologicals 38 (4) (2010) 449–458.
[211] National Institute of Standards and Technology, Development of Standard Reference Materials for Vitamins B_6 and B_{12} in Human Serum [updated February 27, 2015]. Available from: <http://www.nist.gov/mml/analytical/organic/vitsb6andb12inserum.cfm>.
[212] R.L. Bailey, R. Carmel, R. Green, C.M. Pfeiffer, M.E. Cogswell, J.D. Osterloh, et al., Monitoring of vitamin B-12 nutritional status in the United States by using plasma methylmalonic acid and serum vitamin B-12, Am. J. Clin. Nutr. 94 (2) (2011) 552–561.
[213] J. Brady, L. Wilson, L. McGregor, E. Valente, L. Orning, Active B_{12}: a rapid, automated assay for holotranscobalamin on the Abbott AxSYM analyzer, Clin. Chem. 54 (3) (2008) 567–573.

[214] R. Carmel, Subclinical cobalamin deficiency, Curr. Opin. Gastroenterol 28 (2) (2012) 151–158.
[215] E. Nexo, E. Hoffmann-Lucke, Holotranscobalamin, a marker of vitamin B-12 status: analytical aspects and clinical utility, Am. J. Clin. Nutr. 94 (1) (2011) 359s–365ss.
[216] R.L. Bailey, R.A. Durazo-Arvizu, R. Carmel, R. Green, C.M. Pfeiffer, C.T. Sempos, et al., Modeling a methylmalonic acid-derived change point for serum vitamin B-12 for adults in NHANES, Am. J. Clin. Nutr. 98 (2) (2013) 460–467.
[217] NIH, US National Library of Medicine, MedlinePlus. Available from: <https://www.nlm.nih.gov/medlineplus/>.
[218] Institute of Medicine Standing Committee on the Scientific Evaluation of Dietary Reference I, Its Panel on Folate OBV, Choline, The National Academies Collection: Reports Funded by National Institutes of Health. Dietary Reference Intakes for Thiamine, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin, and Choline, National Academies Press, National Academy of Sciences, Washington, DC, 1998.
[219] J.K. Blusztajn, T.J. Mellott, Choline nutrition programs brain development via DNA and histone methylation, Cent. Nerv. Syst. Agents Med. Chem. 12 (2) (2012) 82––94.
[220] L.M. Fischer, K.A. daCosta, L. Kwock, P.W. Stewart, T.S. Lu, S.P. Stabler, et al., Sex and menopausal status influence human dietary requirements for the nutrient choline, Am. J. Clin. Nutr. 85 (5) (2007) 1275–1285.
[221] F.L. Wang, D.R. Haubrich, A simple, sensitive, and specific assay for free choline in plasma, Anal. Biochem. 63 (1) (1975) 195–201.
[222] M. Takayama, S. Itoh, T. Nagasaki, I. Tanimizu, A new enzymatic method for determination of serum choline-containing phospholipids, Clin. Chim. Acta 79 (1) (1977) 93–98.
[223] Y. Ozarda Ilcol, G. Uncu, I.H. Ulus, Free and phospholipid-bound choline concentrations in serum during pregnancy, after delivery and in newborns, Arch. Physiol. Biochem. 110 (5) (2002) 393–399.
[224] K.E. McMahon, P.M. Farrell, Measurement of free choline concentrations in maternal and neonatal blood by micropyrolysis gas chromatography, Clin. Chim. Acta 149 (1) (1985) 1–12.
[225] S.H. Kirsch, W. Herrmann, Y. Rabagny, R. Obeid, Quantification of acetylcholine, choline, betaine, and dimethylglycine in human plasma and urine using stable-isotope dilution ultra performance liquid chromatography-tandem mass spectrometry, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 878 (32) (2010) 3338–3344.
[226] K.J. Brunst, R.O. Wright, K. DiGioia, M.B. Enlow, H. Fernandez, R.J. Wright, et al., Racial/ethnic and sociodemographic factors associated with micronutrient intakes and inadequacies among pregnant women in an urban US population, Public Health Nutr. 17 (9) (2014) 1960–1970.
[227] C.L. Linster, E. Van Schaftingen, Vitamin C. Biosynthesis, recycling and degradation in mammals, FEBS J. 274 (1) (2007) 1–22.
[228] C.L. Linster, S.G. Clarke, L-Ascorbate biosynthesis in higher plants: the role of VTC2, Trends Plant Sci. 13 (11) (2008) 567–573.
[229] J. Mandl, A. Szarka, G. Banhegyi, Vitamin C: update on physiology and pharmacology, Br. J. Pharmacol 157 (7) (2009) 1097–1110.
[230] C.S.S.F. Johnston, R.B. Rucker, Ascorbic acid, in: R.B. Rucker, Sj, D.B. McCormick, L.J. Machlin (Eds.), Handbook of Vitamins, third ed., Marcel Dekker Inc, New York, NY, 2001, pp. 529–554.
[231] A. Mosdol, B. Erens, E.J. Brunner, Estimated prevalence and predictors of vitamin C deficiency within UK's low-income population, J. Public Health (Oxf) 30 (4) (2008) 456–460.
[232] S.T. Omaye, Safety of megavitamin therapy, Adv. Exp. Med. Biol. 177 (1984) 169–203.
[233] D.G. Meyers, P.A. Maloley, D. Weeks, Safety of antioxidant vitamins, Arch. Intern. Med. 156 (9) (1996) 925–935.
[234] J. Blanchard, T.N. Tozer, M. Rowland, Pharmacokinetic perspectives on megadoses of ascorbic acid, Am. J. Clin. Nutr. 66 (5) (1997) 1165–1171.
[235] M. Lucock, Z. Yates, L. Boyd, C. Naylor, J.H. Choi, X. Ng, et al., Vitamin C-related nutrient-nutrient and nutrient-gene interactions that modify folate status, Eur. J. Nutr. 52 (2) (2013) 569–582.
[236] Institute of Medicine, in: N.A. Press (Ed.), Vitamin C, The National Academies Press, Washington, DC, 2000, pp. 95–182.
[237] C.J. Bates, K.S. Jones, L.J. Bluck, Stable isotope-labelled vitamin C as a probe for vitamin C absorption by human subjects, Br. J. Nutr. 91 (5) (2004) 699–705.
[238] A. Karlsen, R. Blomhoff, T.E. Gundersen, Stability of whole blood and plasma ascorbic acid, Eur. J. Clin. Nutr. 61 (10) (2007) 1233–1236.
[239] A. Zinellu, S. Sotgia, L. Deiana, C. Carru, Pre-analytical factors affecting ascorbic and uric acid quantification in human plasma, J. Biochem. Biophys. Methods 67 (2–3) (2006) 95–105.
[240] R. Ferin, M.L. Pavao, J. Baptista, Rapid, sensitive and simultaneous determination of ascorbic and uric acids in human plasma by ion-exclusion HPLC-UV, Clin. Biochem. 46 (7–8) (2013) 665–669.
[241] X. Li, A.A. Franke, Fast HPLC–ECD analysis of ascorbic acid, dehydroascorbic acid and uric acid, J. Chromatogr. B 877 (10) (2009) 853–856.
[242] M.A. Kall, C. Andersen, Improved method for simultaneous determination of ascorbic acid and dehydroascorbic acid, isoascorbic acid and dehydroisoascorbic acid in food and biological samples, J. Chromatogr. B Biomed. Sci. Appl. 730 (1) (1999) 101–111.
[243] H.E. Sauberlich, S.M. Wood, T. Tamura, L.E. Freeberg, Influence of dietary intakes of erythorbic acid on plasma vitamin C analyses, Am. J. Clin. Nutr. 54 (6 Suppl) (1991) 1319s–1322ss.

Self-assessment questions

1. Which of the following are lipid-soluble vitamins? 1) vitamin A; 2) vitamin B12; 3) vitamin D; 4) folate; and 5) vitamin K.
a. 1, 2, 4, 5
b. 1, 3, 5,
c. 2, 3, 4, 5
d. 1, 3, 5
e. All

2. Which statements are true regarding lipid-soluble vitamins? They are 1) absorbed by passive diffusion in all circumstances; 2) absorbed by specific carriers at physiological concentrations; or 3) absorbed by passive diffusion at pharmacological doses?
   a. 1
   b. 2
   c. 3
   d. All

3. Xerophthalmia is due to 1) vitamin D deficiency; 2) folate deficiency; 3) vitamin K deficiency; and 4) vitamin A deficiency.
   a. 1
   b. 2
   c. 3
   d. 4

4. Which of the following statement is false?
   a. The Endocrine Society defines vitamin D deficiency as $25OHD_3$ serum concentrations ≤ 50 nmol/L;
   b. Vitamin E at very high doses may lead to harmful effects such as decreasing blood coagulating capacity, and disrupting granulocyte and phagocyte anti infective function;
   c. Proposed references values for circulating vitamin E lie between 14 and 20 μM;
   d. Patients with cystic fibrosis present with a higher prevalence of vitamin K deficiency.

5. Which of the following statements are true?
   a. The water-soluble vitamin B complex consists of eight vitamins: thiamin, riboflavin, niacin, pantothenic acid, pyridoxine and pyridoxamine, biotin, folic acid and cyanocobalamin;
   b. The small intestine and the pancreas absorb dietary and supplement water-soluble vitamins;
   c. The colon absorbs water-soluble vitamins produced by the microbiota;
   d. Plasma TPP concentration is recognized as being an adequate indicator of body stores.
      1. a, b, c
      2. a, c, d
      3. b, c, d
      4. a, c
      5. all

6. Which of the following statements are true?
   a. Pantothenic acid nutritional status is best reflected by whole blood or erythrocyte concentrations;
   b. Pantothenic acid, in cells containing mitochondria, is present in majority as coenzyme-A;
   c. Pyridoxal-5′-phosphate (PLP) is the principal active cofactor for amino acid transferases, decarboxylases and dehydratases;
   d. Restricted intake, intestinal malabsorption conditions, and chronic alcoholism are the most frequent reasons for PLP deficiency;
   e. Megaloblastic anemia, defined by RBC macrocytosis and abnormal nuclear maturation, is distinguishable from that caused by vitamin B_{12} deficiency.
      1. a, b, c, d
      2. a, b, d, e
      3. b, c, d, e
      4. a, c, d, e
      5. all

Answers

1. b
2. b & c
3. d
4. c
5. 4
6. 1

Chapter 47

Trace elements: functions and assessment of status through laboratory testing

Edgard Delvin[1,2] and Emile Levy[1,3]

[1]Division of Gastroenterology, Hepatology and Nutrition, CHU Sainte Justine Research Centre, Montreal, QC, Canada, [2]Department of Biochemistry, Faculty of Medicine, University of Montreal, Montreal, QC, Canada, [3]Department of Nutrition, Faculty of Medicine, University of Montreal, Montreal, QC, Canada

Learning objectives

After reading this chapter, the reader will be able to:

- Recognize clinically important trace elements.
- Describe situations in which the assessment of trace element status is necessary to prevent deterioration in health status of individuals and/or populations.
- Apply trace element testing to situations requiring the investigation of their status utilizing methods available to clinical and research laboratories.
- Discuss the methodological approaches that are used to measure trace elements.

Introduction

In the context of human nutrition, elements can be classified as bulk elements (calcium, magnesium, phosphorus, sulfur, sodium, potassium, and chloride, the last three being the main electrolytes), and trace elements, needed in minute quantities for the proper growth and development, and present in parts per billion (ppb) in the circulation [1−4]. Care must be taken in distinguishing trace elements originating from environmental contamination of public health concerns (lead, cadmium, cobalt, and mercury) that have toxic effects, even at low concentrations, and functional trace elements (iron, iodine, zinc, selenium, copper, manganese, cobalt, molybdenum, boron, and fluorine) that participate actively in various biological processes. Deficiency and excess in these elements lead to impairment of health status, examples being given in Table 47.1. For the purposes of this chapter, we will review, with varying emphasis, the 10 trace elements that are essential in humans, of which iron, iodine, zinc, selenium, copper, and manganese are the most important. Chromium perceived as a trace element and as a contaminant having toxic properties will also be addressed.

Iron

Iron (Fe), a transition metal with a high redox potential, ubiquitously distributed in biological systems, exists in solution as ferrous (Fe^{2+}), and ferric ions (Fe^{3+}). The majority of Fe is present in hemoglobin, myoglobin, cytochromes, and flavoproteins involved in oxygen transport and the mitochondrial oxidoreduction pathway, and in other enzymes such as catalase and peroxidase. Fe's high redox potential is a double-edged sword, as it can damage cellular components such as fatty acids, proteins, and nucleic acids by an excessive free-radical production in adverse conditions. This characteristic is shared with other trace elements such as copper (Cu) and chromium (Cr), with which Fe may interact [6]. However, the ability of Fe to cause oxidative stress is dampened by being bound to carrier proteins and molecules with antioxidant properties. Body iron stores are mainly regulated through the feedback modulation of iron intestinal absorption that involves iron-regulatory proteins and hepcidin gene regulation [7,8]. Studies on human subjects have shown that calcium, irrespective of its form, can inhibit Fe absorption. However, this interference seems to be of short duration, as adaptation mechanisms develop with time [9]. Ascorbic acid increases absorption by maintaining iron in solution but care should be taken as the combination ascorbic acid−Fe leads to the production of reactive oxygen species (ROS) and the creation of an oxidative stress [6].

Iron deficiency is the most frequent micronutrient deficiency worldwide, particularly in infants, children, adolescents, and women [10]. It generally goes unnoticed until body stores are depleted resulting in iron-deficiency anemia (IDA). Acute iron toxicity occurs most frequently in children causing constipation, vomiting, and diarrhea. Chronic overload is rare in the general population and is

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00047-8

TABLE 47.1 Selected illnesses following improper intakes of trace elements.

Element	Deficiency	Excess
Iron	Anemia	Gastrointestinal effects (constipation, nausea, and vomiting) and genetic hemochromatosis
Iodine	Acquired goitrous hypothyroidism (myxedema), fetal iodine deficiency (mental retardation and deaf mutism), and delayed growth	Chronic and subchronic excess: impairment of thyroid function (hypothyroidism with or without goiter), hyperthyroidism, and thyroid gland malignancies
Zinc	Poor prenatal development, growth retardation, impaired nerve conduction, reproductive failure, dermatitis, alopecia, diarrhea, loss of appetite, anemia, susceptibility to infections, and macular degeneration	Chronic exposure: hypocupremia, impaired iron mobilization, anemia, leukopenia, neutropenia, metabolic disorders, and abnormal cardiac function
Selenium	Kashin–Beck disease (musculoskeletal disorder) and Keshan disease (cardiomyopathy)	Chronic exposure: selenosis (hair, nails, skin lesions, neurological disorders, pain, convulsion, and paralysis)
Copper	Acquired anemia, neutropenia, bone abnormalities, stunting, and congenital Menkes syndrome	Rare occurrence, epigastric pain, nausea, vomiting, diarrhea, childhood cirrhosis (First Nations), Wilson Disease, and increased prevalence of hepatoma?
Manganese	Reported only in experimental conditions, coagulopathy, and dermatitis	Occupational exposure: manganism, a neurotoxic condition similar to Parkinson's disease
Cobalt	Not reported in human populations	Cardiomyopathy in heavy beer drinkers (interaction with ethanol), secondary to associated conditions? Hypothyroidism and poiter, decreased iodine uptake
Molybdenum	Not reported in human populations	Low toxicity index, associated with diarrhea and anemia, and hyperuricemia
Boron	Not officially reported in human populations. Associated with Kashin–Beck disease? Impairment of cognitive functions?	Instability, seizures, and gastrointestinal disturbances
Chromium	Observed in patients on long-term parenteral nutrition, impaired glucose tolerance and utilization, weight loss, neuropathy, dyslipidemia, and abnormalities in nitrogen metabolism (associated with other primary conditions)?	Chronic exposure: renal failure, anemia, hemolysis, liver dysfunction. Symptoms reversible when exposed to Cr^{+3}
Fluorine	Not documented	Fluorosis and osteosclerosis

Source: Expert Group on Vitamins and Minerals. Safe upper levels for vitamins and minerals. 2003. Report No.: 1–904026-11-7.

essentially due to autosomal recessive hereditary hemochromatosis in which the hemochromatosis protein gene (High Fe), a member of the major histocompatiblity complex Class 1 family, is inactivated, thereby causing a dysregulation of Fe uptake [11].

Assessment of iron status

Hemoglobin is the most commonly used parameter in the evaluation of the iron status due to its simplicity and accessibility. However, when used alone, it lacks specificity as it is influenced by vitamin B_{12} and folate status, as well as by infections and hemoglobinopathies. Modern flow cytometry allows accurate and precise measurement of red blood cell morphometry [12]. Mean corpuscular volume reduction (microcytosis) is a reliable index that develops late in severe iron deficiency and when thalassemia and anemia of chronic disease (ACD) have been ruled out. Anisochromia and anisocytosis are also observed in IDA but are not specific to this condition [13].

Biochemical tests to investigate iron deficiency include total serum iron, total iron binding capacity (TIBC), transferrin, soluble transferrin receptor (sTfR), serum ferritin, and erythrocyte zinc protoporphyrin. Their individual value and limitations have been reviewed in the scope of the differential diagnosis of IDA and ACD [14]. In summary, the plasma iron (PI)/TIBC ratio reflecting the transferrin saturation [15] is easily obtained from PI and TIBC routinely analyzed spectrophotometrically on automated platforms. In the presence of iron deficiency, transferrin synthesis is upregulated to maximize iron delivery to target cells, resulting in low transferrin saturation. However, inflammatory and protein-energy wasting conditions may lead to transferrin downregulation that could yield a normal or even high PI/TIBC ratio even in the presence of true iron deficiency state, thereby masking an impairment of iron delivery to hematopoietic cells. Direct measurement of transferrin has been proposed as a better biomarker, and automated immunochemical methods have been developed [16,17] but are not extensively used in clinical laboratories. An LC-MS/MS method using synthetic transferrin peptides has recently been described

TABLE 47.2 Sample reference intervals for serum trace element concentrations.

	Subjects	Age (Y)	Ctry	Methods	Median	GM	IQR	Range	95thile	Mean	SD	Ref.
Fe	141M/F	21–87	NW	ICP-AES	21.4			8.3–41.6				[124]
Fe	119M	19 ± 1	BR	Roche/902					8.4–28.8			[125]
Fe	1400F[a]	27.0 ± 4.5	CN	ICP-MS	50.6				21.5–101.9			[126]
Zn	141M/F	21–87	NW	ICP-AES	13.3			10.4–18.4				[124]
Zn	1400F[a]	27.0 ± 4.5	CN	ICP-MS	24.7				17.3–37.1			[126]
Zn	110M/65F	18-74	BR	ICP-MS		24.6	21.0–28.2					[127]
						23.4	20.0–27.0					
Zn	46M/151F	>67	CZ	FAAS		196.6		100–404.7				[128]
Zn	1016M/F	>70	SE	ICP-MS	11.1		10.1–12.2			11.2	1.75	[129]
Se	141M/F	21–87	NW	ICP-AES	1.26			0.89–2.04				[124]
Se	1400F[a]	27.0 ± 4.5	CN	ICP-MS	2.12				1.17–3.28			[126]
Se	46M/151F	>67	CZ	ICP-MS		2.17		1.07–6.17				[128]
Cu	141M/F	21–87	NW	ICP-AES	17.1			7.7–33.0				[124]
Cu	1400F[a]	27.0 ± 4.5	CN	ICP-MS	67.4				36.9–125			[126]
Cu	110M 65F	18-74	BR	ICP-MS		30.7	25.4–36.9					[127]
						48.3	35.4–62.6					
Cu	46M/151F	>67	CZ	ICP-MS		32.4		22.5–53.4				[128]
Cu	1016M/F	>70	SE	ICP-MS	14.9		13.1–16.8			15.1	2.92	[129]
Mn[b]	141M/F	21–87	NW	ICP-AES	2.55			<55.70				[124]
Mn[b]	1400F[a]	27.0 ± 4.5	CN	ICP-MS	96				48–336			[126]
Mn[b]	110Mv65F	18-74	BR	ICP-MS		21.2	15.8–32.7					[127]
						22.4	14.7–33.5					
Mn[b]	46M/151F	>67	CZ	ICP-MS		436		24–3704				[128]
Mn[b]	1016M/F	>70	SE	ICP-MS	21.1		14.8–35.9			28.4	21.0	[129]
Co[b]	141M/F	21–87	NW	ICP-AES	<10[c]			<130				[124]
Co[b]	1400F[a]	27.0 ± 4.5	CN	ICP-MS	12.5				2.1–41.8			[126]
Co[b]	110M	18-74	BR	ICP-MS		5.6	4.2–8.8					[127]
	65F					5.9	4.1–9.4					
Co[b]	1016M/F	>70	SE	ICP-MS	1.7		1.3–2.4			2.44	5.36	[129]
Mo[b]	1400F[a]	27.0 ± 4.5	CN	ICP-MS	42.8				28.5–85.7			[126]
Mo[b]	1016M/F	>70	SE	ICP-MS	11.3		9.0–14.7			15.2	30.9	[129]
B	69M/72F	21–87	NW	ICP-AES	1.50			<3.74				[124]
Cr[b]	141M/F	21–87	NW	ICP-AES	<70[c]			<350				[124]
Cr[b]	1400F[a]	27.0 ± 4.5	CN	ICP-MS	34.9				9.2–178.6			[126]
Cr[b]	110Mv65F	18-74	BR	ICP-MS	0.085	81.3	70.0–102.9					[127]
					0.078	81.9	67.7–106.4					
Cr[b]	1016M/F	>70	SE	ICP-MS	9.7		6.3–15.8			14.9	20.7	[129]
F[b]	53F	20–22	JN	ISE				194–984		383	158	[130]

95thile, 95th percentile; *Age*, ages are expressed either as range or mean ± SD; *BR*, Brazil; *CN*, China; *Ctry*, Country; *CZ*, Czech Republic; *FAAS*, flame atomic absorption spectrometry; *GM*, geometric mean; *ICP-AES*, inductively coupled plasma atomic emission spectrometry; *ICP-MS*, Inductively coupled plasma mass spectrometry; *IQR*, Interquartile range; *ISE*, ion-selective electrode; *JN*, Japan; *Md*, median; *NW*, Norway; *R*, range; *Ref.*, reference; *SD*, standard deviation; *SE*, Sweden.

[a] *Pregnant women. Values are given as μmol/L unless otherwise indicated.*

[b] *nmol/L.*

[c] *Value given is the limit of detection.*

for the measurement of serum transferrin [18]. It could become the basis for a reference method that would allow improving the accuracy of immunological methods.

sTfR measurement, a more recently described assessor of iron status, typically distinguishes between IDA and ACD, and is particularly useful in identifying IDA in patients affected by concurrent inflammatory diseases in which ferritin, an acute phase protein, is increased. Measurement of sTfR is of particular clinical interest, as the concentration increases rapidly when iron stores are deficient, falls before other biomarkers upon normalization of the iron status, and is not influenced by acute phase responses [19]. ELISAs and automated immunoturbidimetric and chemiluminescent sTfR assays have been described. They are, however, not harmonized, thereby impeding the definition of a common diagnostic threshold and decreasing its clinical sensitivity and specificity [19–21]. The establishment, in 2010 of a Reference Reagent to standardize immunoassays for sTfR by the WHO Expert Committee on Biological Standardization should help in solving this issue [22].

Serum ferritin was thought to be a better biomarker to detect early iron deficiency. However, being an acute phase reactant, its serum concentration is influenced by inflammatory conditions and may be elevated by chronic alcohol use, thereby limiting its utility in the diagnosis of IDA [23–25]. Furthermore, the heterogeneity of ferritin isoforms, the panoply of antibodies used in immunoassays, as well as the absence of an internationally recognized method have impaired the clinical interpretation of the results. This problem has partially been solved by the use of the third International Standard to assure the traceability of the different automated chemiluminescent and electrochemiluminescent assay standardization [26].

Erythrocyte zinc protoporphyrin, a product of abnormal heme synthesis, is particularly valuable in the primary screening of uncomplicated iron deficiency (iron-deficiency erythropoiesis) and, together with sTfR, for monitoring iron supplement therapy, when lead toxicity (rare nowadays) is ruled out [27]. Low costs and ability to measure the erythrocyte zinc protoporphyrin/heme ratio directly on a drop of blood with a point-of-care hematofluorimeter are the major advantages of this test. The difficulty in automating the assay, however, limits its use by central clinical laboratories. In last recourse, bone marrow may be examined to assess the amount of hemosiderin in the reticulum cells. However, this technique is invasive, semiquantitative, and observer-dependent and cannot be used routinely in clinical pathology laboratories.

Iodine

Hypothyroidism can be classified as goitrous, the most visible sign of endemic iodine deficiency and of goitrogenic foods consumption, and nongoitrous, the congenital form of hypothyroidism [28]. Despite massive iodination of salt, the 2009 UNICEF survey revealed that more than 200 million globally distributed school-age children suffered from iodine deficiency [29]. Iodine (I_2) deficiency disorders are major public health concerns, as they lead to brain damage and mental retardation, particularly in children and to goiter in adults. Furthermore, severe in utero I_2 deficiency retards fetal growth and brain development. The recognition of the universality of iodine deficiency highlights the need to develop and apply new strategies to establish and maintain iodine dietary status monitoring programs.

Dietary I_2, absorbed in the small intestine, reaches the circulation in the form of iodide (I^-), from which it is cleared by the Na^+/I^- symporter (NIS) into the thyroid gland, where it is concentrated in the follicular cell as I_2 for thyroid hormone biosynthesis [28]. It is also transported into other tissues such as mammary gland to provide I^- to the neonate or cleared by the kidney. In euthyroid adults, the thyroid gland contains approximately 70%–80% of the body I_2, and I_2 uptake varies greatly according to the nutritional status [30]. Pituitary thyrotropin [thyroid stimulating hormone (TSH)] stimulates all stages of thyroid iodine metabolism, including NIS-mediated iodine transport into the thyroid gland, and the synthesis and secretion of thyroxin (T_4) and triiodothyronine (T_3). Approximately 90% of the unnecessary I^- is excreted in urine, and the remaining 10% in the feces and perspiration. Iatrogenic iodine toxicity may be due to exposure to radiological contrast media, drugs, and iodine-containing sterilizing solutions. It usually results in high plasma iodine concentrations, derangement of thyroid hormone synthesis accompanied by brassy taste, increased salivation, gastric irritation, and acne-like lesions.

Assessment of iodine status

The global nature of iodine deficiency warrants the need to maintain and strengthen I_2 nutritional surveillance programs. Their efficient management and evaluation require good quality data on urine I_2 concentrations. The 2001 WHO shift from goiter prevalence as the indicator of I_2 deficiency to the spot-urine median I_2 excretion level, the principal marker of recent I_2 intake, was a pivotal decision that dramatically improved the accuracy of the prevalence data [31]. However, as individual I_2 intakes vary from day to day, interpretation of spot-urine I_2 values warrants some caution [32]. Although I_2 concentrations measured on 24-hour collections are more accurate, such an approach is more demanding, particularly in population studies.

The alternative, the age- and sex-adjusted iodine/creatinine ratio, is more accurate than the casual reporting based on concentration per liter, but may lead to

overestimation in the presence of low protein intake (low urinary creatinine) and underestimation in conditions of high creatinine excretion [33]. Knudsen et al. [33] have reported a median (5%–95%iles) 24-hour iodine urinary excretion 143 (75–297) μg/day in a small group of apparently healthy adults.

The spectrophotometric measurement of Sandell–Kolthoff reduction reaction based on iodide-catalyzed reaction between arsenic (III) and cerium (IV), stopped by the addition of diphenylamine-4-sulfonic acid, is the method mostly used to measure urinary I^- [34]. This method has been adapted to a microplate robotic system, allowing the measurement of large numbers of samples [35]. A more technological advanced method using the inductively coupled plasma mass-spectrometry method (ICP-MS) has also been described [36]. Its analytical qualities and stability make this approach particularly valuable for long-term monitoring of populations. ICP-MS can also be coupled to multielement analyses, offering an additional advantage.

Serum TSH and serum thyroglobulin (Tg) are complementary "functional" biochemical indicators of the iodine status, reflecting intermediate time response to the nutritional status. Their use is, however, limited by the cost and technology needed in the perspective of a national screening program, particularly in developing countries in which the iodine deficiency is endemic. Whereas TSH is a sensitive indicator of iodine status in the newborn period, it is relatively insensitive in older children and adult, as values could remain within reference intervals especially in the presence of a borderline iodine deficiency. Furthermore, data must be interpreted with some caution, since normative values are age-, physiological state-, and analytical platform-dependent [37,38]. Tg is a storage protein providing the substance for thyroxin synthesis. In normal conditions, serum Tg concentration is low. It increases when the thyroid becomes hyperplastic, is correlated with urinary I_2, and declines rapidly upon iodine repletion, making it a sensitive indicator of iodine homeostasis [39,40]. Automated immunochemical assay methods for measuring Tg in serum or dried blood spots have been published [41–43]. Although they have a high throughput potential, their use in epidemiological surveys is limited by cost and because of the potential presence of Tg heterophilic antibodies that can artificially decrease the Tg concentration, which would require a reflex test after neutralizing the antibodies.

Zinc

Zinc is an essential Group IIB posttransition element of public health importance. Based on food availability data and on the 2002 WHO report, zinc deficiency affects approximately 1/3 of the world population [44]. Also based on the 2004 Canadian Community Health Survey and the 2003–2006 NHANES survey, the standardized prevalence of inadequate zinc intakes in Canadian and American food-secure and food-insecure households was 13% and 29%, and 12% and 17%, respectively [45]. Zinc deficiency may also be observed in the presence of high phytate content diets due to poor intestinal absorption. Zinc and copper mutually interfere with the gastrointestinal (GI) uptake of the other, thus potentially leading to zinc imbalance in the presence of high copper-containing diets. Similarly, zinc and iron compete for absorption.

Zinc is an essential constituent of zinc-dependent endopeptidases that play key roles in the remodeling of extracellular matrices and of nuclear transcription factors that regulate the transcription of genes involved in metabolic and immune processes, and cell life cycle [46–49]. Zinc deficiency is characterized by poor prenatal development and intrauterine growth retardation, short stature, hypogonadism, impaired immune function, skin disorders, cognitive dysfunction, anorexia, and loss of taste and smell, and is responsible worldwide for approximately 16% of lower respiratory tract infections, 18% of malaria, and 10% of diarrheal diseases [44]. Intracellular Zn homeostasis mainly depends on two species of Zn transporters (ZnTs): Zrt- and Irt-like proteins, known as solute carrier family 39A, ZnTs, known as SLC30A proteins, and metallothioneins (MTs) [50].

Rare genetic loss of expression and activity of ZnTs has been involved in the pathogenesis of acrodermatitis enteropathica, Spondylocheiro dysplastic Ehlers–Danlos syndrome, transient neonatal zinc deficiency, Type II diabetes, and hepatic cirrhosis [51]. Acute zinc salt toxicity, accidental or iatrogenic, is accompanied by variety of symptoms, including abdominal pain, nausea vomiting, lethargy, anemia, and dizziness. Chronic overexposure to zinc can result in secondary hypocupremia, by the inhibition of intestinal copper absorption, which symptoms include impaired iron mobilization, hematological disorders, decreased ROS detoxification superoxide dismutase (SOD) activity, decreased ceruloplasmin, decreased cytochrome c oxidase, and impairment of lipid metabolism [5].

Assessment of zinc status

Despite its limitations in detecting marginal deficiency, its responsiveness to acute and chronic inflammation, and its circadian fluctuations, serum zinc concentration remains the most commonly used biomarker for zinc status. Although urinary zinc excretion declines in parallel to zinc body stores, it is an insensitive measure of zinc status. Several precautions must be observed during phlebotomy, as zinc contamination can occur if zinc-containing ointments are used. Hemolysis must be avoided, since the erythrocyte zinc concentration is greater than serum or

plasma. The type of collecting tube is also of importance; glass "trace metal" vacuum tubes (glass) are most frequently used. Although the utilization of plastic tubes has not been advised because of potential environmental contamination, Boeynaems et al. [52] polyethylene terephthalate tubes were suitable alternatives to glass tubes. Serum/plasma zinc is customarily measured by either flame atomic absorption spectroscopy or ICP-MS [53,54]. MT monocyte messenger RNA, which is downregulated in zinc deficiency and upregulated with zinc supplementation, has been proposed as a surrogate "functional marker" of zinc deficiency. Its use is, however, limited as it is not within reach of all clinical laboratories, lacks reference intervals, and is influenced by inflammatory processes. Furthermore, one study has demonstrated its poor predictive value for poor zinc nutritional status [55].

Selenium

Selenium is an essential trace element for the production of selenoproteins that act either as structural proteins or as enzymes. The mechanisms involved in their synthesis are unique in that selenium is incorporated through the opal UGA stop codon, usually read as a translation stop signal, and is there perceived as a sense selenocysteine (Sec) codon. This phenomenon requires a conserved Sec insertion sequence, located in the 3′-untranslated region, which then recruits a binding protein (SBP2) to serve as a scaffold for mobilizing a Sec elongation factor [56]. Selenoenzymes are catalysts for a number of metabolic pathways related to antioxidant balance, thyroid hormones, and immunity [28,57]. The mechanisms involved in the intestinal selenium absorption are not completely understood. However, an in vitro transport model using human Caco-2 cells shows that selenium is absorbed from the apical pole as selenomethionine and methyl selenocysteine through saturable systems [58]. Hepatic production of methylated selenium compounds and their urinary excretion regulate body selenium content. Selenium is primarily transported in the plasma to the organs via Sepp1, a apolipoprotein E receptor-2 [57].

Selenium deficiency is thought to be a risk factor for several chronic conditions linked to oxidative stress and inflammation involving the GI tract, muscles, and the neurological system [59,60] among others. Selenium has also been associated with the endemic Keshan disease, a congestive cardiomyopathy, resulting from an interaction between its deficiency and the presence of a mutant strain of Coxsackie virus, as well as Kashin−Beck disease, an endemic osteochondropathy thought to result from the combined selenium deficiency, and the presence of mycotoxins in grain high levels of humic acid in drinking water [61,62]. Like other micronutrients, selenium deficiency has also been suspected in patients receiving total parenteral nutrition, in which cardiomyopathy, muscle pain, and muscle weakness are resolved by selenium supplementation. The range between selenium deficiency and toxicity is narrow. Frequent symptoms associated with acute toxicity include diarrhea, fatigue, alopecia, joint pain, nail discoloration or brittleness, and nausea. Longer exposition to excessive selenium peripheral neuropathy and dermatitis is observed [63,64].

Assessment of selenium status

Circulating selenium increasing quickly with intake, plasma, or serum selenium concentration is the most frequently used assessor of short-term selenium dietary consumption. Several techniques for the measurement of serum/plasma/urine selenium have been described. Although older methods utilized fluorometry, the present-day methods are based on atomic absorption spectrophotometry (AAS) with Zeeman correction or ICP-MS without digestion [65].

The study of Meplan et al. [66] opened the possibility of using the circulating Selenoprotein P (SePP) concentration and its polymorphism to evaluate longer term selenium status. They showed, by Western blot analysis, that the serum concentration of SePP-1 depended on the genotype and the selenium concentration and that a 6-week selenium supplementation produced a shift in the proportion of two SePP-1 isoforms. An immunoassay for one specific polymorphic SePP-1 could possibly be developed and used as a "functional" biomarker of selenium status. Plasma glutathione peroxidase-3, erythrocyte glutathione peroxidase-1, and blood cell selenoperoxidase activities, all dependent on selenium are also potential biomarkers. However, their specificity is lacking, as they may be modulated by numerous health conditions.

Copper

Copper (Cu) is a transition metal, widely distributed in nature with cupric ion (Cu^{2+}) being the most important oxidative state. It is ubiquitously distributed in cells and tissues with highest concentrations in brain and liver. In essence, the body's copper reserve is almost totally bound to enzyme prosthetic groups or to chaperone proteins that assure its delivery to specific target proteins such as superoxide dismustase-1 and Cu(1) P_{1B}-type ATPases [67]. Copper is absorbed from the intestinal lumen through the highly conserved apical enterocyte membrane copper-transporter family (Human Ctr1 transporter) and picked up by specific chaperones [68]. Copper is a cofactor of enzymes involved in several metabolic processes required for the development of tissue structures. It is also important in ROS control through ceruloplasmin (ferroxidase I), cytochrome c oxidase, SOD, tyrosinase, lysyl

oxidase, amine oxidases, dopamine-beta-monooxygenase, and peptidylglycine monooxygenase reactions. Because of its central role in biological processes, copper deficiency leads to hematological, musculoskeletal, dermatological, neurological, immunological, and developmental problems. Copper deficiency is fortunately rare, but has been observed in malnourished infants, premature and low birth weight infants fed cow's milk, and patients on copper-free total parenteral nutrition [69].

Copper, similarly to iron, is a double-edged sword as it can also be a toxic element, as it is observed in hepatic, neurodegenerative, and cardiovascular diseases. Its destructive potential is attributed to its ability, as is iron, to take part in Fenton-like reactions yielding extremely harmful ROS [70]. Fortunately, the efficiency of the copper homeostatic control allows a low incidence of chronic copper toxicosis. Copper poisoning can nevertheless develop under certain conditions. The clinical features include vomiting, diarrhea, GI bleeding, hemolysis, oliguria, hematuria, seizures, coma, and eventually death if not treated [71].

Two well-known genetic diseases involve copper transport. The first, Menke's disease or kinky-hair syndrome, could be perceived as a copper pseudodeficiency syndrome due to an ubiquitous (except liver) inactivating mutation of the copper-transporter P-type ATPase (P_{1B} ATP7A) preventing the copper, taken up at the enterocyte luminal membrane of being secreted into the portal circulation, thereby reducing copper bioavailability and decreasing copper-dependent enzyme activities. The second is Wilson's disease picturing a copper toxicity syndrome resulting from an inactivating mutation of the copper-transporter P-type ATPase (P_{1B} ATP7B), preventing copper from being transported to the Golgi apparatus for secretion, and thus accumulating in the hepatocytes. Of importance in both situations, serum/plasma copper and ceruloplasmin, common biomarkers of copper status, are decreased, as they are in nutritional copper deprivation. In Wilson's disease, urinary copper excretion and liver content are increased [72].

Assessment of copper status

Total Serum or plasma copper ($>80\%$ ceruloplasmin-bound), and ceruloplasmin are both widely utilized as biomarkers of the copper status. Their validity, however, may not as reliable as originally ascertained. Although they are decreased in the presence of severe copper deficiency, their concentrations may vary in the cases of mild deficiency. Furthermore, both demonstrate variability in their circulating concentrations related to age, sex, and pregnancy, and are increased by conditions such as inflammatory or infectious processes, cancer, and estrogen therapy, which can hide copper deficiency [73].

Other markers of copper status have been evaluated including 24-hour-urine copper and several cuproenzymes including leukocyte cytochrome c oxidase, erythrocyte SOD, serum diamine oxidase, and plasma peptidylglycine α-amidating monooxygenase. However, the lack of standardization, the absence of normative values, and high interindividual variability limit their use for diagnostic purposes. The recent observation of the increased abundance of the peripheral mononuclear cell copper chaperone for superoxide dismutase (CCS) in malnourished children supports the theory that CCS may be a copper biomarker [74]. Future research may result in the addition of these potential markers to the more traditional plasma copper and ceruloplasmin measures. In addition, as contamination is of particular concern in trace metal analysis; plasma copper measurement must be performed using trace element-free vacuum tubes. The two methods of choice for measuring this element are with AAS with Zeeman correction and ICP-MS [54,75].

Manganese

Manganese, a transition element for which $+2$ oxidative state is the most stable, is a cofactor for a variety of enzymes including oxidoreductases, dismutases, transferases, hydrolases, lyases, isomerases, and ligases [76]. It also modulates the affinity of some carbohydrate-binding lectins and transmembrane receptors (integrins) for their respective ligands [77–79]. Manganese and other trace elements (copper, zinc, fluorine, magnesium, iron, and boron) play a role in bone metabolism, being a cofactor for enzymes synthesizing bone matrix glycoproteins. Although manganese could not be singled out, a double-blind placebo-controlled supplementation study in osteoporotic women that included this element plus calcium, zinc and copper resulted in a net increase in spinal bone mineral density [80]. These data argue in favor of an active role for manganese and the other nutrients in preserving bone density and preventing osteoporosis. The divalent metal ion transporter-1 transports only a small percentage of dietary manganese through the intestinal apical membrane [81]. The integrity of the enterohepatic cycle is essential as the absorbed manganese is excreted very rapidly into the gut via concentration gradient into the bile duct [82]; hence, the risk of manganese toxicity is higher in neonates, in whom this cycle is immature, and also in obstructive liver disease.

No clinical deficiency has been reported in healthy individuals with poor dietary intakes. In a depletion study, seven adult volunteers who were fed a low manganese-containing diet for 35 days developed dermatological lesions accompanied by a decline in serum cholesterol, which resolved quickly upon repletion. Manganese toxicity is a well-recognized occupational risk for miners who are likely to inhale manganese dust if not properly protected. Manganese overload affects the extrapyramidal

motor system in particular, leading to lesions and symptoms close to those of Parkinson's disease and to psychiatric conditions [79].

Assessment of manganese status

The assessment of manganese nutritional status remains problematic. Although plasma/serum manganese concentrations are often cited as reflecting dietary manganese intake, they tend to react only to wide intake variations. Urinary excretion follows the same pattern when usual amounts of manganese are consumed and seems to be responsive only to severe depletion [79,83]. Hence, further studies are needed to evaluate their usefulness as indicators of manganese status. A recently proposed cohort study to be conducted among apprentice welders to determine if blood could be a useful biomarker for manganese status in the context of an occupational health surveillance program will help solving the problem [84].

Liver arginase activity has been shown to be depressed in manganese-deficient animals and lymphocyte Mn-SOD activity increased in supplemented women [79]. These functional tests could in the future potentially be biomarkers for manganese status, providing that simplified methods are developed. As with the other trace elements, serum and urinary manganese can be measured by ICP-MS in samples collected in trace element-free tubes, and in whole blood or erythrocytes by graphite furnace AAS [54,75].

Cobalt

Cobalt, a transition metal for which oxidative states +2 and +3 are the most common, is the key constituent of cobalamin, which is its reservoir in the body (see the section on Vitamin B_{12} in Chapter 46, Vitamins and Nutritional Testing). Cobalt toxicity is not frequent and is more likely to result from occupational exposure in industries where its concentration in ambient air is high. The use of cobalt–chromium hard-metal alloys in arthroplasty is also a potential source of internal cobalt exposure, as slow corrosion and deterioration create nanometric-size metal debris that are systemically disseminated [85]. Finally, the use of cobalt salts by high-level athletes for enhancing their aerobic performance is another potential cause of cobalt contamination [86,87]. These improper intakes of inorganic cobalt salts can cause severe GI tract, thyroid, heart, and sensory system damages [88,89].

Assessment of cobalt status

It is generally accepted that plasma or serum cobalt includes both free cobalt (Co^{2+}) and protein-bound soluble cobalt ions, which reflects the recent exposure, and that red blood cell cobalt content reflects the long-term exposure to Co^{2+} [86,90]. A urinary and fecal excretion study demonstrated that Co^{2+} was primarily eliminated in urine the first 2–8 days after administration, showing that urine could serve for biomonitoring a recent exposure [91]. ICP-MS is the method of choice to measure total cobalt content in biological samples [92]. Trace element-free tubes with strict sampling procedures should be used to avoid environmental contamination.

Molybdenum

Molybdenum (Mo), an essential trace element and crucial for the survival of animals, shuttles between the oxidation states Mo^{+2} and Mo^{+6}. Four Mo-dependent enzymes are known in humans: sulfite oxidase, xanthine oxidase, aldehyde oxidase, and mitochondrial amidoxime reductase, each containing a pterin-based molybdenum prosthetic group acting as cofactor [93]. The mechanisms involved in molybdenum absorption from the human intestinal lumen and in its transcellular transport and release into the lymphatic system have not yet been characterized. However, presence of the molybdate transporter CrMoT2 in human embryonic kidney cell line HEK-293T may allow to further our understanding of the transport process [94].

To date, molybdenum deficiency due to an inadequate intake has not been reported in healthy individuals. Adult patients and preterm infants on long-term parenteral nutrition without supplementation may, however, develop a deficiency. Multielement deficiency usually accompanies this nowadays-rare deficiency. Human molybdenum cofactor deficiency is a rare recessive inborn error of metabolism (incidence below 1:100,000), which results in the loss of activity of sulfite oxidase, xanthine dehydrogenase, and aldehyde oxidase. Symptoms that develop soon after birth include seizures, progressive encephalopathy and other progressive neurological damage, and, in most cases to, early childhood death [95]. There is limited evidence for molybdenum toxicity in humans. Hyperuricemia and arthralgia have been associated with increased dietary molybdenum intake and with increased serum molybdenum concentration. However, no other study has confirmed this observation [93]. The rapid urinary excretion of molybdenum could possibly explain its low toxicity. However, occupational exposure, by inhalation, to molybdenum containing dusts, such as molybdenum trioxide (MoO_3), has been associated with induced subclinical alveolitis [96]. Finally, Meeker et al. [97] have provided evidence for molybdenum being a male reproductive toxicant.

Assessment of molybdenum status

In plasma, molybdenum circulates bound specifically to α2-macroglobulin. In terms of excretion, experimental data from healthy volunteers show that, in repletion states,

>70% of the absorbed amount is excreted via urine and <8% in the feces [98]. As with other trace minerals, molybdenum is measured by ICP-MS and inductively coupled plasma optical emission spectroscopy [54].

Boron

Boron is a metalloid element found in nature and foods, principally nuts, fruits, and green vegetables, mainly as boric acid and salt. It is readily absorbed from the human gut through mechanisms that are yet to be defined. In addition to passive diffusion, in vitro results suggest the existence of a boron-specific membrane transporter in murine monocyte cell lines [99] and aquaporin channel-mediated transport in plants [100]. The human counterpart of these pathways have, however, not been confirmed.

The exact function of boron is elusive. There is, however, some suggestion that it is beneficial to bone growth, musculoskeletal health, and central nervous system function, among other systems. Several lines of evidence suggest that boron activity manifests through the formation of boroesters, biomolecules such as S-adenosylmethionine, diadenosine phosphates, as well as nicotinamide adenine dinucleotide, phosphoinositides, glycoproteins, and glycolipids [101]. The high bone boron content supports the hypothesis that it may be an important element in bone growth and function by contributing to the metabolism and fixation of calcium [102].

Boron deficiency is rare or poorly documented in human population studies. However, it has been suggested that it could be, together with germanium and selenium, a contributing factor to Kashin–Beck disease [62,103]. Boron compounds have a low order of toxicity and accidental lethal poisoning is nowadays extremely rare. Acute boron poisoning generally presents with GI (nausea, persistent vomiting, abdominal pain, and diarrhea) and neurological symptoms (headache, irritability, delirium, seizures, and occasionally coma and death) [104]. The most frequently cited consequence of chronic exposure to boron relates to male reproductive toxicity. The few epidemiologic studies that have examined human reproductive health related to environmental or workplace borates diverged in their conclusions, with some reporting no adverse effects and some reporting sterility [105]; this question remains pending.

Assessment of boron status

Reports on boron homeostasis are scarce. Sutherland et al. [106] demonstrated that urinary boron excretion changed rapidly with modifications in boron intake, suggesting that the kidney was the primary site for regulating its homeostasis and that urine was thus a sensitive indicator of its intake. The preferred methods for the analysis of boron are inductively coupled plasma atomic emission spectroscopy and more commonly ICP-MS [1].

Chromium

Chromium is a metallic element that exists in a variety of oxidation states, of which trivalent (Cr^{3+}) and hexavalent (Cr^{6+}) ions are the biologically most important, with (Cr^{6+}) containing compounds being considered toxic and mutagenic [107]. Grain products, fruits, and vegetables provide most of the dietary chromium [108] at variable levels. Chromium deficiency in the general population, although suggested, is poorly documented lack of sensitive nutritional indicators. This might be a false problem, since the status of chromium as an essential element, and even as a pharmacologic agent, is being challenged and its alleged beneficial effect on insulin sensitivity not having been conclusively demonstrated in type 2 diabetic patients [109–111]. Chromium deficiency has been reported in total parenteral nutrition patients who manifested impaired glucose tolerance, hyperglycemia, weight loss, and neuropathy.

Chromium is absorbed by passive diffusion when taken orally and its uptake from the intestine is increased when the paracellular channels are disturbed (increased permeability) such as in type 2 diabetes [109,112]. Once absorbed and transported to the portal system, chromium is bound to transferrin, the major iron blood transport protein [113]. Chromium toxicity is valence-dependent, with Cr^{+6} having a much higher degree of toxicity than Cr^{+3}. However, given that the latter is poorly absorbed, its toxicity is low. The adverse effects of exposure to high Cr^{+6} containing environment include renal failure, genotoxicity, hepatic dysfunction, and rhabdomyolysis [79]. Irritation of skin, lungs, and GI tract and perforation of the nasal septum have also reported in the cases of suspected chromium overload.

Assessment of chromium status

Reliable biomarkers of dietary chromium and chromium status are lacking, as normal plasma levels are close to the limit of detection of currently available assays. However, elevated plasma chromium may be a good indicator of recent dietary of or excessive exposure to chromium. Graphite furnace AAS with Zeeman correction and ICP-MS are the two common methods for measuring chromium in serum or urine [92,114].

Fluorine

Fluorine, a strongly reactive halide, is widely available in the environment as the ion fluoride (F^-). Primary sources are tea, water (particularly supplemented), and dental products; it can also be absorbed as a contaminant in pesticides (cryolite and sulfuryl fluoride) [115]. Fluoride is not

considered an essential nutrient, because its deficiency has not been defined, and clinical signs of deprivation have not been reported [116]. Given that the basis for assessing a low fluoride status is nonexistent, fluoride deficiency has not been defined. Dental fluorosis expressed as porous enamel is the most overt sign of fluoride toxicity. Excess fluoride consumption may also increase the number of osteoblasts and generate osteosclerosis [117]. The Expert Committee on fluoride in drinking water of the National research Council has identified young children, patients with diabetes and chronic renal disease, elderly persons, hypersensitive individuals, pregnant women or lactating mothers, and individuals deficient in specific micronutrients such as calcium, magnesium, iodine, and selenium at risk of adverse effects for fluoride excess [115]. Experiments conducted in confluent-differentiated Caco-2 cells indicate that the paracellular pathway leads the bidirectional fluoride transport (absorption and secretion). However, the participation of a F^-/H^+ cotransporter or a F^-/OH^- antiporter cannot be excluded [118].

Assessment of fluorine status

Serum, saliva, and urine have all been used to evaluate recent fluoride exposure. Urine being the prime vehicle for fluoride excretion, 24-hour collections are particularly useful in assessing current fluoride exposure for population studies [119]. Serum fluoride concentration has also proven to be useful in monitoring bone turnover in postmenopausal osteoporosis treatment with alendronate [120]. Nail clippings is an unusual biologic material that proved its utility in evaluating long-term exposure to fluoride ions. Since fluoride enters fingernails at the growth end and reaches the distal end by approximately 3 months, its concentration in the clipping reflects the average fluoride intake concentration during the period when the clipping was obtained [121]. Potentiometry using ion-selective electrodes is the most frequently used technology to measure fluoride ions [122]. Gas chromatography-mass spectrometry after derivatization with 2-(bromomethyl)-naphthalene has also been described [123]. This method is, however, technologically more demanding and not practical for most first-line clinical laboratories. Using the ion-selective electrode approach, Zohoori et al. [122] have reported urinary fluoride excretion values ranging from 0.03 to 0.476 mg/days. Table 47.2 provides a sample of reference intervals for the serum concentration of each trace element covered in this chapter [124–130].

References

[1] J.P. Goulle, L. Mahieu, J. Castermant, N. Neveu, L. Bonneau, G. Laine, et al., Metal and metalloid multi-elementary ICP-MS validation in whole blood, plasma, urine and hair. Reference values, Forensic Sci. Int. 153 (1) (2005) 39–44.

[2] J.P. Goulle, P. Le Roux, M. Castanet, L. Mahieu, S. Guyet-Job, M. Guerbet, Metallic profile of whole blood and plasma in a series of 99 healthy children, J. Anal. Toxicol. 39 (9) (2015) 707–713.

[3] M. Hambidge, Biomarkers of trace mineral intake and status, J. Nutr. 133 (Suppl 3) (2003) 948s–955ss.

[4] E. Barany, I.A. Bergdahl, L.E. Bratteby, T. Lundh, G. Samuelson, A. Schutz, et al., Trace element levels in whole blood and serum from Swedish adolescents, Sci. Total Environ. 286 (1–3) (2002) 129–141.

[5] Expert Group on Vitamins and Minerals. Safe Upper Levels for Vitamins and Minerals, 2003. Report No.: 1–904026-11-7.

[6] K. Jomova, M. Valko, Advances in metal-induced oxidative stress and human disease, Toxicology 283 (2–3) (2011) 65–87.

[7] T. Ganz, Hepcidin and iron regulation, 10 years later, Blood 117 (17) (2011) 4425–4433.

[8] L.C. Kuhn, Iron regulatory proteins and their role in controlling iron metabolism, Metallomics 7 (2) (2015) 232–243.

[9] B. Lonnerdal, Calcium and iron absorption--mechanisms and public health relevance, Int. J. Vitam. Nutr. Res. 80 (4–5) (2010) 293–299.

[10] M.B. Zimmermann, R.F. Hurrell, Nutritional iron deficiency, Lancet 370 (9586) (2007) 511–520.

[11] W.J. Griffiths, Review article: the genetic basis of haemochromatosis, Aliment. Pharmacol. Ther. 26 (3) (2007) 331–342.

[12] E. Urrechaga, L. Borque, J.F. Escanero, Biomarkers of hypochromia: the contemporary assessment of iron status and erythropoiesis, BioMed Res. Int. 2013 (2013) 603786.

[13] J. Ford, Red blood cell morphology, Int. J. Lab. Hematol. 35 (3) (2013) 351–357.

[14] S.F. Clark, Iron deficiency anemia, Nutr. Clin. Pract. 23 (2) (2008) 128–141.

[15] G. Siek, J. Lawlor, D. Pelczar, M. Sane, J. Musto, Direct serum total iron-binding capacity assay suitable for automated analyzers, Clin. Chem. 48 (1) (2002) 161–166.

[16] R. Gambino, E. Desvarieux, M. Orth, H. Matan, T. Ackattupathil, E. Lijoi, et al., The relation between chemically measured total iron-binding capacity concentrations and immunologically measured transferrin concentrations in human serum, Clin. Chem. 43 (12) (1997) 2408–2412.

[17] R. Castel, M.G. Tax, J. Droogendijk, M.P. Leers, R. Beukers, M. D. Levin, et al., The transferrin/log(ferritin) ratio: a new tool for the diagnosis of iron deficiency anemia, Clin. Chem. Lab. Med. 50 (8) (2012) 1343–1349.

[18] T.A. Zimmerman, M. Wang, M.S. Lowenthal, I.V. Turko, K.W. Phinney, Quantification of transferrin in human serum using both QconCAT and synthetic internal standards, Anal. Chem. 85 (21) (2013) 10362–10368.

[19] F. Braga, I. Infusino, A. Dolci, M. Panteghini, Soluble transferrin receptor in complicated anemia, Clin. Chim. Acta 431 (2014) 143–147.

[20] C.M. Pfeiffer, J.D. Cook, Z. Mei, M.E. Cogswell, A.C. Looker, D.A. Lacher, Evaluation of an automated soluble transferrin receptor (sTfR) assay on the Roche Hitachi analyzer and its comparison to two ELISA assays, Clin. Chim. Acta 382 (1–2) (2007) 112–116.

[21] I. Infusino, F. Braga, A. Dolci, M. Panteghini, Soluble transferrin receptor (sTfR) and sTfR/log ferritin index for the diagnosis of iron-deficiency anemia. A meta-analysis, Am. J. Clin. Pathol. 138 (5) (2012) 642–649.

[22] S.J. Thorpe, A. Heath, G. Sharp, J. Cook, R. Ellis, M.A. Worwood, WHO reference reagent for the Serum Transferrin Receptor (sTfR): international collaborative study to evaluate a recombinant soluble transferrin receptor preparation, Clin. Chem. Lab. Med. 48 (6) (2010) 815–820.

[23] M. Yoneda, Y. Nozaki, H. Endo, H. Mawatari, H. Iida, K. Fujita, et al., Serum ferritin is a clinical biomarker in Japanese patients with nonalcoholic steatohepatitis (NASH) independent of HFE gene mutation, Dig. Dis. Sci. 55 (3) (2010) 808–814.

[24] S.K. Seo, B.H. Yun, S.J. Chon, Y.J. Lee, E.J. Han, J.H. Park, et al., Association of serum ferritin levels with metabolic syndrome and subclinical coronary atherosclerosis in postmenopausal Korean women, Clin. Chim. Acta 438 (2015) 62–66.

[25] S.K. Park, W.J. Choi, C.M. Oh, M.G. Kim, W.T. Ham, J.M. Choi, et al., Clinical significance of serum ferritin level as an independent predictor of insulin resistance in Korean men, Diabetes Res. Clin. Pract. 107 (1) (2015) 187–193.

[26] S. Blackmore, M. Hamilton, A. Lee, M. Worwood, M. Brierley, A. Heath, et al., Automated immunoassay methods for ferritin: recovery studies to assess traceability to an international standard, Clin. Chem. Lab. Med. 46 (10) (2008) 1450–1457.

[27] J.D. Cook, Diagnosis and management of iron-deficiency anaemia, Best. Pract. Res. Clin. Haematol 18 (2) (2005) 319–332.

[28] A. Djemli, G. Van Vliet, E.E. Delvin, Congenital hypothyroidism: from paracelsus to molecular diagnosis, Clin. Biochem. 39 (5) (2006) 511–518.

[29] M. Andersson, B. de Benoist, L. Rogers, Epidemiology of iodine deficiency: salt iodisation and iodine status, Best. Pract. Res. Clin. Endocrinol. Metab. 24 (1) (2010) 1–11.

[30] M. Li, C.J. Eastman, The changing epidemiology of iodine deficiency, Nat. Rev. Endocrinol. 8 (7) (2012) 434–440.

[31] World Health Organization UNCsFICftCoID, Disorders. Assessment of Iodine Deficiency Disorders and Monitoring Their Elimination. A Guide for Programme Managers, World Health Organization, Geneva, 2001.

[32] S. Andersen, J. Karmisholt, K.M. Pedersen, P. Laurberg, Reliability of studies of iodine intake and recommendations for number of samples in groups and in individuals, Br. J. Nutr. 99 (4) (2008) 813–818.

[33] N. Knudsen, E. Christiansen, M. Brandt-Christensen, B. Nygaard, H. Perrild, Age- and sex-adjusted iodine/creatinine ratio. A new standard in epidemiological surveys? Evaluation of three different estimates of iodine excretion based on casual urine samples and comparison to 24h values, Eur. J. Clin. Nutr. 54 (4) (2000) 361–363.

[34] O.M. Trokhimenko, V.N. Zaitsev, Kinetic determination of iodide by the Sandell–Kolthoff reaction using diphenylamine-4-sulfonic acid, J. Anal. Chem. 59 (5) (2004) 491–494.

[35] A. Mina, E.J. Favaloro, J. Koutts, A robust method for testing urinary iodine using a microtitre robotic system, J. Trace Elem. Med. Biol. 25 (4) (2011) 213–217.

[36] P. Macours, J.C. Aubry, B. Hauquier, J.M. Boeynaems, S. Goldman, R. Moreno-Reyes, Determination of urinary iodine by inductively coupled plasma mass spectrometry, J. Trace Elem. Med. Biol. 22 (2) (2008) 162–165.

[37] D. Springer, V. Bartos, T. Zima, Reference intervals for thyroid markers in early pregnancy determined by 7 different analytical systems, Scand. J. Clin. Lab. Invest. 74 (2) (2014) 95–101.

[38] K.L. Coene, A.Y. Demir, M.A. Broeren, P. Verschuure, E.G. Lentjes, A.K. Boer, Subclinical hypothyroidism: a 'laboratory-induced' condition? Eur. J. Endocrinol. 173 (4) (2015) 499–505.

[39] A. Djemli, M. Fillion, J. Belgoudi, R. Lambert, E.E. Delvin, W. Schneider, et al., Twenty years later: a reevaluation of the contribution of plasma thyroglobulin to the diagnosis of thyroid dysgenesis in infants with congenital hypothyroidism, Clin. Biochem. 37 (9) (2004) 818–822.

[40] N. Knudsen, I. Bulow, T. Jorgensen, H. Perrild, L. Ovesen, P. Laurberg, Serum Tg--a sensitive marker of thyroid abnormalities and iodine deficiency in epidemiological studies, J. Clin. Endocrinol. Metab. 86 (8) (2001) 3599–3603.

[41] M.B. Zimmermann, B. de Benoist, S. Corigliano, P.L. Jooste, L. Molinari, K. Moosa, et al., Assessment of iodine status using dried blood spot thyroglobulin: development of reference material and establishment of an international reference range in iodine-sufficient children, J. Clin. Endocrinol. Metab. 91 (12) (2006) 4881–4887.

[42] C. Roger, K. Chikh, V. Raverot, F. Claustrat, F. Borson-Chazot, C. Bournaud-Salinas, et al., New-generation thyroglobulin assay: performance and implications for follow-up of differentiated thyroid carcinoma, Ann. Endocrinol 75 (4) (2014) 227–231.

[43] C. Spencer, J. LoPresti, S. Fatemi, How sensitive (second-generation) thyroglobulin measurement is changing paradigms for monitoring patients with differentiated thyroid cancer, in the absence or presence of thyroglobulin autoantibodies, Curr. Opin. Endocrinol Diabetes Obes. 21 (5) (2014) 394–404.

[44] A. Rodgers, Quantifying selected major risks to health, in: C.L.A. Murray (Ed.), TheWorld Heath Report, World Health Organization, Geneva, Switzerland, 2002, pp. 47–97.

[45] S.I. Kirkpatrick, K.W. Dodd, R. Parsons, C. Ng, D. Garriguet, V. Tarasuk, Household food insecurity is a stronger marker of adequacy of nutrient intakes among Canadian compared to American youth and adults, J. Nutr. 145 (7) (2015) 1596–1603.

[46] A. Kuhad, P. Singh, K. Chopra, Matrix metalloproteinases: potential therapeutic target for diabetic neuropathic pain, Expert. Opin. Ther Targets 19 (2) (2015) 177–185.

[47] N. Spellmon, J. Holcomb, L. Trescott, N. Sirinupong, Z. Yang, Structure and function of SET and MYND domain-containing proteins, Int. J. Mol. Sci. 16 (1) (2015) 1406–1428.

[48] L. Huang, Zinc and its transporters, pancreatic beta-cells, and insulin metabolism, Vitam. Hormones 95 (2014) 365–390.

[49] P. Bonaventura, G. Benedetti, F. Albarede, P. Miossec, Zinc and its role in immunity and inflammation, Autoimmun. Rev. 14 (4) (2015) 277–285.

[50] T. Takagishi, T. Hara, T. Fukada, Recent advances in the role of SLC39A/ZIP zinc transporters in vivo, Int. J. Mol. Sci. 18 (12) (2017).

[51] T. Kambe, A. Hashimoto, S. Fujimoto, Current understanding of ZIP and ZnT zinc transporters in human health and diseases, Cell. Mol. Life Sci 71 (17) (2014) 3281–3295.

[52] J.M. Boeynaems, A. De Leener, B. Dessars, H.R. Villa-Lobos, J. C. Aubry, F. Cotton, et al., Evaluation of a new generation of plastic evacuated blood-collection tubes in clinical chemistry, therapeutic drug monitoring, hormone and trace metal analysis, Clin. Chem. Lab. Med. 42 (1) (2004) 67–71.

[53] M.R. Islam, M.U. Ahmed, S.A. Mitu, M.S. Islam, G.K. Rahman, M.M. Qusar, et al., Comparative analysis of serum zinc, copper,

manganese, iron, calcium, and magnesium level and complexity of interelement relations in generalized anxiety disorder patients, Biol. Trace Elem. Res. 154 (1) (2013) 21–27.
[54] J.M. Harrington, D.J. Young, A.S. Essader, S.J. Sumner, K.E. Levine, Analysis of human serum and whole blood for mineral content by ICP-MS and ICP-OES: development of a mineralomics method, Biol. Trace Elem. Res. 160 (1) (2014) 132–142.
[55] C.S. Kwon, A.M. Kountouri, C. Mayer, M.J. Gordon, I.S. Kwun, J.H. Beattie, Mononuclear cell metallothionein mRNA levels in human subjects with poor zinc nutrition, Br. J. Nutr. 97 (2) (2007) 247–254.
[56] M.J. Berry, Insights into the hierarchy of selenium incorporation, Nat. Genet. 37 (11) (2005) 1162–1163.
[57] S. Kurokawa, M.J. Berry, Selenium. Role of the essential metalloid in health, Met. Ions Life Sci. 13 (2013) 499–534.
[58] C. Thiry, A. Ruttens, L. Pussemier, Y.J. Schneider, An in vitro investigation of species-dependent intestinal transport of selenium and the impact of this process on selenium bioavailability, Br. J. Nutr. 109 (12) (2013) 2126–2134.
[59] S.J. Fairweather-Tait, Y. Bao, M.R. Broadley, R. Collings, D. Ford, J.E. Hesketh, et al., Selenium in human health and disease, Antioxid. Redox Signal. 14 (7) (2011) 1337–1383.
[60] B. Speckmann, H. Steinbrenner, Selenium and selenoproteins in inflammatory bowel diseases and experimental colitis, Inflamm. Bowel Dis. 20 (6) (2014) 1110–1119.
[61] J. Chen, An original discovery: selenium deficiency and Keshan disease (an endemic heart disease), Asia Pac. J. Clin. Nutr. 21 (3) (2012) 320–326.
[62] Y. Yao, F. Pei, P. Kang, Selenium, iodine, and the relation with Kashin-Beck disease, Nutrtion 27 (11–12) (2011) 1095–1100.
[63] J.K. MacFarquhar, D.L. Broussard, P. Melstrom, R. Hutchinson, A. Wolkin, C. Martin, et al., Acute selenium toxicity associated with a dietary supplement, Arch. Intern. Med. 170 (3) (2010) 256–261.
[64] K.L. Nuttall, Evaluating selenium poisoning, Ann. Clin. Lab. Sci. 36 (4) (2006) 409–420.
[65] H.C. Wang, H.W. Peng, M.S. Kuo, Determination of beryllium and selenium in human urine and of selenium in human serum by graphite-furnace atomic absorption spectrophotometry, Anal. Sci. 17 (4) (2001) 527–532.
[66] C. Meplan, F. Nicol, B.T. Burtle, L.K. Crosley, J.R. Arthur, J.C. Mathers, et al., Relative abundance of selenoprotein P isoforms in human plasma depends on genotype, se intake, and cancer status, Antioxid. Redox Signal. 11 (11) (2009) 2631–2640.
[67] A.K. Boal, A.C. Rosenzweig, Structural biology of copper trafficking, Chem. Rev. 109 (10) (2009) 4760–4779.
[68] P.A. Sharp, Ctr1 and its role in body copper homeostasis, Int. J. Biochem. Cell Biol. 35 (3) (2003) 288–291.
[69] K. Schumann, H.G. Classen, H.H. Dieter, J. Konig, G. Multhaup, M. Rukgauer, et al., Hohenheim consensus workshop: copper, Eur. J. Clin. Nutr. 56 (6) (2002) 469–483.
[70] L.M. Gaetke, H.S. Chow-Johnson, C.K. Chow, Copper: toxicological relevance and mechanisms, Arch. Toxicol. 88 (11) (2014) 1929–1938.
[71] F. Woimant, J.M. Trocello, Disorders of heavy metals, Handbook of Clinical Neurology 120 (2014) 851–864.
[72] H. Kodama, C. Fujisawa, W. Bhadhprasit, Inherited copper transport disorders: biochemical mechanisms, diagnosis, and treatment, Curr. Drug. Metab. 13 (3) (2012) 237–250.
[73] M. Olivares, M.A. Mendez, P.A. Astudillo, F. Pizarro, Present situation of biomarkers for copper status, Am. J. Clin. Nutr. 88 (3) (2008) 859s–862ss.
[74] M. Arredondo, G. Weisstaub, M. Medina, M. Suazo, M. Guzman, M. Araya, Assessing chaperone for Zn, Cu-superoxide dismutase as an indicator of copper deficiency in malnourished children, J. Trace Elem. Med. Biol 28 (1) (2014) 23–27.
[75] H.W. Zhao, J. Lin, X.B. Wang, X. Cheng, J.Y. Wang, B.L. Hu, et al., Assessing plasma levels of selenium, copper, iron and zinc in patients of Parkinson's disease, PLoS One 8 (12) (2013) e83060.
[76] J.D. Crowley, D.A. Traynor, D.C. Weatherburn, Enzymes and proteins containing manganese: an overview, Met. Ions Biol. Syst. 37 (2000) 209–278.
[77] G. Sun, H. Zhao, B. Kalyanaraman, N.M. Dahms, Identification of residues essential for carbohydrate recognition and cation dependence of the 46-kDa mannose 6-phosphate receptor, Glycobiology 15 (11) (2005) 1136–1149.
[78] P.H. Weinreb, S. Li, S.X. Gao, T. Liu, R.B. Pepinsky, J.A. Caravella, et al., Dynamic structural changes are observed upon collagen and metal ion binding to the integrin alpha1 I domain, J. Biol. Chem. 287 (39) (2012) 32897–32912.
[79] Medicine Io, Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc, The National Academies Press, Washington, DC, 2001, p. 800 p.
[80] L. Strause, P. Saltman, K.T. Smith, M. Bracker, M.B. Andon, Spinal bone loss in postmenopausal women supplemented with calcium and trace minerals, J. Nutr. 124 (7) (1994) 1060–1064.
[81] A. Shawki, P.B. Knight, B.D. Maliken, E.J. Niespodzany, B. Mackenzie, H(+)-coupled divalent metal-ion transporter-1: functional properties, physiological roles and therapeutics, Curr. Top. Membr. 70 (2012) 169–214.
[82] D.S. Avila, R.L. Puntel, M. Aschner, Manganese in health and disease, Met. Ions Life Sci. 13 (2013) 199–227.
[83] J.L. Greger, C.D. Davis, J.W. Suttie, B.J. Lyle, Intake, serum concentrations, and urinary excretion of manganese by adult males, Am. J. Clin. Nutr. 51 (3) (1990) 457–461.
[84] M.G. Baker, C.D. Simpson, B. Stover, L. Sheppard, H. Checkoway, B.A. Racette, et al., Blood manganese as an exposure biomarker: state of the evidence, J. Occup. Environ. Hyg. 11 (4) (2014) 210–217.
[85] M. Oldenburg, R. Wegner, X. Baur, Severe cobalt intoxication due to prosthesis wear in repeated total hip arthroplasty, J. Arthroplast 24 (5) (2009). 825.e15–20.
[86] L.O. Simonsen, H. Harbak, P. Bennekou, Cobalt metabolism and toxicology--a brief update, Sci. Total Environ. 432 (2012) 210–215.
[87] G. Lippi, M. Franchini, G.C. Guidi, Blood doping by cobalt. Should we measure cobalt athletes? J. Occup. Med. Toxicol 1 (2006) 18.
[88] B. Ebert, W. Jelkmann, Intolerability of cobalt salt as erythropoietic agent, Drug. Test. Anal. 6 (3) (2014) 185–189.
[89] C.A. Lim, J. Khan, E. Chelva, R. Khan, T. Unsworth-Smith, The effect of cobalt on the human eye, Doc. Ophthalmol. Adv. Ophthalmol. 130 (1) (2015) 43–48.
[90] L.O. Simonsen, A.M. Brown, H. Harbak, B.I. Kristensen, P. Bennekou, Cobalt uptake and binding in human red blood cells, Blood Cells Mol. Dis. 46 (4) (2011) 266–276.

[91] R.W. Leggett, The biokinetics of inorganic cobalt in the human body, Sci. Total Environ. 389 (2–3) (2008) 259–269.
[92] H.J. Choi, S.J. Lim, Y.S. Park, S.Y. Lee, Simple and robust ICP-MS method for simultaneous determination of serum Co and Cr in routine clinical practice, Clin. Chim. Acta 439 (2015) 91–96.
[93] Institute of Medicine, Molybdenum. Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc, The National Academies Press, Washington, DC, 2001, pp. 420–441.
[94] Y. Nakanishi, S. Iida, H. Ueoka-Nakanishi, T. Niimi, R. Tomioka, M. Maeshima, Exploring dynamics of molybdate in living animal cells by a genetically encoded FRET nanosensor, PLoS One 8 (3) (2013) e58175.
[95] J. Reiss, J.L. Johnson, Mutations in the molybdenum cofactor biosynthetic genes MOCS1, MOCS2, and GEPH, Hum. Mutat. 21 (6) (2003) 569–576.
[96] H.C. Ott, C. Prior, M. Herold, M. Riha, G. Laufer, G. Ott, Respiratory symptoms and bronchoalveolar lavage abnormalities in molybdenum exposed workers, Wien. Klinische Wochenschr. 116 (Suppl 1) (2004) 25–30.
[97] J.D. Meeker, M.G. Rossano, B. Protas, M.P. Diamond, E. Puscheck, D. Daly, et al., Cadmium, lead, and other metals in relation to semen quality: human evidence for molybdenum as a male reproductive toxicant, Environ. Health Perspect. 116 (11) (2008) 1473–1479.
[98] J.R. Turnlund, W.R. Keyes, G.L. Peiffer, G. Chiang, Molybdenum absorption, excretion, and retention studied with stable isotopes in young men during depletion and repletion, Am. J. Clin. Nutr. 61 (5) (1995) 1102–1109.
[99] N.V. Ralston, C.D. Hunt, Transmembrane partitioning of boron and other elements in RAW 264.7 and HL60 cell cultures, Biol. Trace Elem. Res. 98 (2) (2004) 181–191.
[100] E. Bastias, N. Fernandez-Garcia, M. Carvajal, Aquaporin functionality in roots of Zea mays in relation to the interactive effects of boron and salinity, Plant. Biol. 6 (4) (2004) 415–421.
[101] F.H. Nielsen, Update on human health effects of boron, J. Trace Elem. Med. Biol 28 (4) (2014) 383–387.
[102] I. Zofkova, P. Nemcikova, P. Matucha, Trace elements and bone health, Clin. Chem. Lab. Med 51 (8) (2013) 1555–1561.
[103] X. Peng, Z. Lingxia, G.N. Schrauzer, G. Xiong, Selenium, boron, and germanium deficiency in the etiology of Kashin-Beck disease, Biol. Trace Elem. Res. 77 (3) (2000) 193–197.
[104] M.V. Pahl, B.D. Culver, N.D. Vaziri, Boron and the kidney, J. Ren. Nutr 15 (4) (2005) 362–370.
[105] W.A. Robbins, L. Xun, J. Jia, N. Kennedy, D.A. Elashoff, L. Ping, Chronic boron exposure and human semen parameters, Reprod. Toxicol 29 (2) (2010) 184–190.
[106] B. Sutherland, P. Strong, J.C. King, Determining human dietary requirements for boron, Biol. Trace Elem. Res. 66 (1–3) (1998) 193–204.
[107] P. Koedrith, H. Kim, J.I. Weon, Y.R. Seo, Toxicogenomic approaches for understanding molecular mechanisms of heavy metal mutagenicity and carcinogenicity, Int. J. Hyg. Env. Health 216 (5) (2013) 587–598.
[108] R.A. Anderson, N.A. Bryden, M.M. Polansky, Dietary chromium intake. Freely chosen diets, institutional diet, and individual foods, Biol. Trace Elem. Res. 32 (1992) 117–121.
[109] J.B. Vincent, Chromium: is it essential, pharmacologically relevant, or toxic, Met. Ions Life Sci. 13 (2013) 171–198.
[110] S. Lewicki, R. Zdanowski, M. Krzyzowska, A. Lewicka, B. Debski, M. Niemcewicz, et al., The role of Chromium III in the organism and its possible use in diabetes and obesity treatment, Ann. Agric. Environ. Med. 21 (2) (2014) 331–335.
[111] C.H. Bailey, Improved meta-analytic methods show no effect of chromium supplements on fasting glucose, Biol. Trace Elem. Res. 157 (1) (2014) 1–8.
[112] F. Horton, J. Wright, L. Smith, P.J. Hinton, M.D. Robertson, Increased intestinal permeability to oral chromium (51 Cr) -EDTA in human Type 2 diabetes, Diabet. Med. 31 (5) (2014) 559–563.
[113] G. Deng, K. Wu, A.A. Cruce, M.K. Bowman, J.B. Vincent, Binding of trivalent chromium to serum transferrin is sufficiently rapid to be physiologically relevant, J. Inorg. Biochem. 143 (2015) 48–55.
[114] A.M. Roussel, M. Andriollo-Sanchez, M. Ferry, N.A. Bryden, R. A. Anderson, Food chromium content, dietary chromium intake and related biological variables in French free-living elderly, Br. J. Nutr. 98 (2) (2007) 326–331.
[115] N.R. Council, Fluoride in Drinking Water: A Scientific Review of EPA's Standards, The National Academies Press, Washington, DC, 2006, p. 530.
[116] Intakes SCotSEoDR, IOM Tables. In: Sciences NAo, editor. Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride, National Academy Press, Washington, DC, 1997, pp. 288–313.
[117] M.P. Whyte, W.G. Totty, V.T. Lim, G.M. Whitford, Skeletal fluorosis from instant tea, J. Bone Min. Res. 23 (5) (2008) 759–769.
[118] R.A. Rocha, V. Devesa, D. Velez, In vitro study of intestinal transport of fluoride using the Caco-2 cell line, Food Chem. Toxicol. 55 (2013) 156–163.
[119] A.J. Rugg-Gunn, A.E. Villa, M.R. Buzalaf, Contemporary biological markers of exposure to fluoride, Monogr. Oral. Sci. 22 (2011) 37–51.
[120] H. Sato, K. Tanno, G. Muro-oka, K. Itai, Serum ionic fluoride concentrations are significantly decreased after treatment with alendronate in patients with osteoporosis, Clin. Chim. Acta 412 (23–24) (2011) 2146–2149.
[121] G.M. Whitford, Monitoring fluoride exposure with fingernail clippings, Schweiz. Monatssch. Zahnmed. 115 (8) (2005) 685–689.
[122] F.V. Zohoori, R. Walls, L. Teasdale, D. Landes, I.N. Steen, P. Moynihan, et al., Fractional urinary fluoride excretion of 6–7-year-old children attending schools in low-fluoride and naturally fluoridated areas in the UK, Br. J. Nutr. 109 (10) (2013) 1903–1909.
[123] S.M. Kwon, H.S. Shin, Sensitive determination of fluoride in biological samples by gas chromatography-mass spectrometry after derivatization with 2-(bromomethyl)naphthalene, Anal. Chim. Acta 852 (2014) 162–167.
[124] R. Rahil-Khazen, B.J. Bolann, R.J. Ulvik, Trace element reference values in serum determined by inductively coupled plasma atomic emission spectrometry, Clin. Chem. Lab. Med. 38 (8) (2000) 765–772.
[125] L.A. Nunes, H.Z. Grotto, R. Brenzikofer, D.V. Macedo, Hematological and biochemical markers of iron status in a male, young, physically active population, BioMed. Res. Int. 2014 (2014) 349182.

[126] X. Liu, Y. Zhang, J. Piao, D. Mao, Y. Li, W. Li, et al., Reference values of 14 serum trace elements for pregnant Chinese women: a cross-sectional study in the China nutrition and health survey 2010–2012, Nutrients 9 (3) (2017).
[127] G.H. Rocha, C. Steinbach, J.R. Munhoz, M.A. Madia, J.K. Faria, D. Hoeltgebaum, et al., Trace metal levels in serum and urine of a population in southern Brazil, J. Trace Elem. Med. Biol. 35 (2016) 61–65.
[128] J. Rambouskova, A. Krskova, M. Slavikova, M. Cejchanova, K. Wranova, B. Prochazka, et al., Trace elements in the blood of institutionalized elderly in the Czech Republic, Arch. Gerontol. Geriatr 56 (2) (2013) 389–394.
[129] B. Schultze, P.M. Lind, A. Larsson, L. Lind, Whole blood and serum concentrations of metals in a Swedish population-based sample, Scand. J. Clin. Lab. Invest. 74 (2) (2014) 143–148.
[130] K. Itai, H. Tsunoda, Highly sensitive and rapid method for determination of fluoride ion concentrations in serum and urine using flow injection analysis with a fluoride ion-selective electrode, Clin. Chim. Acta 308 (1–2) (2001) 163–171.

Self-assessment questions

1. Which elements other than Fe share the potential of excessive free-radical production in adverse conditions?
2. Which statements are true regarding the concentration of soluble serum transferrin receptor (sTfR)? Its level (i) increases rapidly when iron stores are deficient; (ii) falls long before additional biomarkers upon normalization of the iron status; and (iii) is influenced by acute phase responses?
   - **a.** i
   - **b.** ii
   - **c.** iii
   - **d.** All
3. What statement is true? Goiter is the sign of (i) endemic iodine deficiency and (ii) congenital hypothyroidism.
   - **a.** 1
   - **b.** 2
   - **c.** both
4. Which of the following statement is false?
   - **a.** Zinc, copper, and selenium are classified as essential elements.
   - **b.** Zinc deficiency can be observed in the presence of high phytate contained in diets.
   - **c.** Selenoenzymes are catalysts for metabolic pathways related to antioxidant balance.
   - **d.** Selenium deficiency is a risk factor for chronic conditions linked to oxidative stress and inflammation.
   - **e.** Measurement of serum selenium is used for assessing long-term selenium intake.
5. Which of the following four statements are true?
   - **a.** Manganese is a cofactor for oxidoreductases, dismutases, and transferases.
   - **b.** Manganese, copper, zinc, and fluoride play important roles in bone metabolism.
   - **c.** Cobalt toxicity is frequently observed in rural environment.
   - **d.** Sulfite oxidase and xanthine oxidase are molybdenum-dependent enzymes.
   - **e.** Plasma chromium reflect recent dietary intake.
   - **1.** a, b, c, d
   - **2.** a, b, c, e
   - **3.** a, b, d, e
   - **4.** b, c, d, e
6. Which of the following statements are true?
   - **a.** Chromium is a metallic element that exists in a variety of oxidation states of which trivalent (Cr^{3+}) and hexavalent (Cr^{6+}) ions are the biologically most important.
   - **b.** The preferred methods for analysis of boron are inductively coupled with plasma atomic emission spectroscopy (ICP-AES) and more commonly inductively coupled with plasma mass spectrometry (ICP-MS).
   - **c.** In plasma, molybdenum circulates as a free element.
   - **d.** Red blood cell cobalt content reflects the long-term exposure to Co^{2+}.

Answers

1. a, b, c, d
2. a, b, d
3. b, c, d
4. a, c, d
5. a, b, c

Chapter 48

Newborn screening and inborn errors of metabolism

Khushbu Patel[1] **and Stephen R. Master**[2]

[1]Division of Pediatric Pathology, Department of Pathology, UT Southwestern Medical Center, Dallas, TX, United States, [2]Division of Laboratory Medicine, Children's Hospital of Philadelphia, Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- List the major groups of inherited metabolic disorders and associated laboratory findings.
- Discuss the role of newborn screening and the methodology used for detecting metabolic diseases.
- Describe commonly used analytical methodologies for evaluation of inborn errors of metabolism.

Introduction to inborn errors of metabolism

Approximately one quarter of pediatric hospital admissions are due to genetic diseases, a large subset of which are a class of metabolic diseases known as inborn errors of metabolism (IEMs). While individually these are classified as rare diseases, collectively their incidence approaches 1:500. Most IEMs are inherited in an autosomal recessive manner and occur due to defects in biochemical pathways involved in metabolism and energy production. They arise from either a loss of function of an enzyme directly involved in a metabolic step or deficiencies in proteins, vitamins, or cofactors required for enzyme stability and/or in vivo activity. The age of onset and severity of symptoms often depends on the type of IEM and other factors such as mutation status and external triggers of metabolic decompensation (e.g., fasting, infection, and fever). While in some cases IEMs are not compatible with life, there are other instances where signs and symptoms of disease may be completely absent even in the presence of a biochemical abnormality.

Newborns affected with an IEM often have a nonspecific clinical presentation or a long asymptomatic period. Most signs and symptoms of IEMs are due to either accumulation of toxic metabolites or metabolite deficiencies that are important for energy production. Clinically, IEMs can be grouped accordingly: (1) disorders that give rise to intoxication; (2) disorders involving energy metabolism; and (3) disorders involving complex molecules (Table 48.1). In disorders of intoxication, metabolites proximal to the metabolic block accumulate to toxic concentrations. Biochemical measurement of these metabolites forms the basis for diagnosis of most IEMs. Deficiencies of energy production involve the Krebs cycle, fatty acid oxidation, or mitochondrial respiration. Disorders involving complex molecules arise from defects in synthesis or catabolism of complex molecules, and include lysosomal and peroxisomal disorders, as well as disorders of intracellular processing and trafficking. These defects lead to the accumulation of metabolites within various cellular compartments.

The laboratory plays a vital role in diagnosis of IEMs, which requires performing the correct test on the most appropriate sample type and properly interpreting test results. IEMs may be diagnosed through routine and specialized tests. Many of the screening tests used in the evaluation of metabolic disease use instrumentation similar to that available in many clinical chemistry laboratories. What distinguishes metabolic testing from routine testing are the compounds analyzed and their correlation to IEMs. Most physicians requesting metabolic testing are not familiar with the metabolic abnormalities seen in various IEMs. For this reason, most metabolic tests require an interpretation in addition to any quantitative results. Test reports should ideally include the result (with age-appropriate reference intervals for quantitative results), a differential diagnosis based on the pertinent abnormal and normal findings, recommendations for further testing,

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00048-X

TABLE 48.1 Categories of inborn errors of metabolism and their associated clinical and biochemical findings.

IEM category	Clinical features	Associated laboratory findings	IEM types
Intoxication	Neurological distress Seizures Lethargy/coma Liver/other organ failure Unusual urine odor/color	Metabolic acidosis High anion gap Hyperammonemia Ketonuria	Organic acidurias Amino acid disorders Urea cycle disorders Galactosemia
Energy deficiency	Cardiomyopathy Hepatomegaly Lethargy/coma Exercise intolerance	Hypoglycemia (nonketotic) Lactic acidosis Ketosis	Fatty acid oxidation defects Glycogen storage diseases
Disorders of complex metabolism	Hydrops fetalis Hepatomegaly Hypotonia Seizures Neurological abnormalities Severe growth and mental retardation		Lysosomal disorders Peroxisomal disorders Congenital disorders of glycosylation

IEM, Inborn errors of metabolism.

information regarding additional recommended testing, and contact information for the person writing the report should questions arise. Results are interpreted most accurately when clinical and relevant laboratory information is included with the test requisition. When results point to a number of IEMs, correlation with the patient's clinical and laboratory findings is necessary to narrow the differential diagnosis and suggest the most appropriate follow-up testing.

Newborn screening

In the United States, most common IEMs are detected through newborn screening programs. Newborn screening is a population-based screening method to identify newborns with metabolic, endocrine, and other disorders for which timely detection and treatment prevents severe complications and manifestations of the disease. A number of criteria must be met for a condition to be included on the screening panel. These criteria are based on the Wilson and Jungner principles developed in 1968 on behalf of the World Health Organization. For a condition to be included on a screening panel, a test with appropriate sensitivity and specificity must be available, and clear benefits through early detection and timely intervention must be demonstrated. Careful consideration must be given to each criterion, weighing its impact on both affected and unaffected individuals.

In 2006 the American College of Medical Genetics recommended a list of 29 core conditions, which was approved by the Secretary Advisory Committee for Heritable Diseases in Newborn and Children (SACHDNC) and formed the Recommended Uniform Screening Panel (RUSP). The SACHDNC periodically updates the RUSP using an evidence-based protocol for proposing and approving a condition. To date, there are 35 core conditions and 26 secondary conditions on the RUSP. Though not all disorders on newborn screening panels are IEMs, a vast majority of these conditions (~80%) are metabolic disorders. Table 48.2 lists the IEMs that are currently included on the RUSP, as well as the estimated incidence of the core conditions. It is important to note that screening is not nationally mandated in the United States; individual state public health departments guide the number of disorders included in the screening, and the methods used for screening.

Methods used in newborn screening programs are geared toward direct measurement or detection of metabolites, measurement of enzyme activity, or DNA analysis. This testing is typically performed at a centralized laboratory. Screening for other conditions, such as critical congenital heart disease and hearing loss, requires functional tests, performed at the birthing facility. The first method used for newborn screening dates back to the 1960s, which started with screening for phenylketonuria (PKU) using a bacterial inhibition assay known as the Guthrie test. Advances in tandem mass spectrometry have now replaced this method and have allowed for expansion of current newborn screening programs. Currently, over 40 metabolic disorders can be screened for from blood spotted on a filter paper card or "Guthrie card" by measuring a combination of amino acids and acylcarnitines with only 2–4 minutes of instrument time.

Newborn screening is recognized as one of the most successful public health programs. Newborn screening

TABLE 48.2 Metabolic conditions included on the Recommended Uniform Screening Panel and their estimated incidence.

	Core conditions	Incidence	Secondary conditions
Organic acid disorders	Propionic acidemia	1:200,000	Malonic acidemia
	Methylmalonic aciduria	1:100,000	Isobutyrylglycinuria
	Isovaleric acidemia	1:80,000	2-Methylbutyrylglycinuria
	3-Methylcrotonyl-CoA carboxylase deficiency	1:50,000	3-Methylglutaconic aciduria
	HMG-CoA lyase deficiency	<1:100,000	2-Methyl-3-hydroxybutyric aciduria
	Holocarboxylase synthase deficiency	<1:200,000	
	ß-Ketothiolase deficiency	<1:200,000	
	Glutaric aciduria type I	1:100,000	
Fatty acid oxidation disorders	Carnitine uptake defect/carnitine transport defect	1:22,000	Short-chain acyl-CoA dehydrogenase deficiency
	Medium-chain acyl-CoA dehydrogenase deficiency	1:15,000	Medium/short-chain L-3-hydroxyacyl-CoA dehydrogenase deficiency
	Very long-chain acyl-CoA dehydrogenase deficiency	1:85,000	Glutaric acidemia type II
	Long-chain L-3 hydroxyacyl-CoA dehydrogenase deficiency	<1:250,000	Medium-chain ketoacyl-CoA thiolase deficiency
	Trifunctional protein deficiency		2,4-Dienoyl-CoA reductase deficiency
			Carnitine palmitoyltransferase type I deficiency
			Carnitine palmitoyltransferase type II deficiency
			Carnitine acylcarnitine translocase deficiency
Amino acid/urea cycle disorders	Argininosuccinic aciduria	1:220,000	Argininemia
	Citrullinemia type I	1:250,000	Citrullinemia type II
	Maple syrup urine disease	1:185,000	Hypermethioninemia
	Homocystinuria	1:450,000	Benign hyperphenylalaninemia
	Classic phenylketonuria	1:10,000	Biopterin defect in cofactor biosynthesis
	Tyrosinemia type I	1:100,000	Biopterin defect in cofactor regeneration
			Tyrosinemia type II
			Tyrosinemia type III
Other IEMs	Biotinidase deficiency	1:100,000	Galactoepimerase deficiency
	Classic galactosemia	1:60,000	Galactokinase deficiency
	Glycogen storage disease	1:20,000	
	Pompe disease	1:40,000	
	Mucopolysaccharidosis type 1	1:100,000	
	X-linked adrenoleukodystrophy	1:17,000	

CoA, Coenzyme A; *IEM*, inborn errors of metabolism; *HMG-CoA*, β-hydroxy β-methylglutaryl-CoA.

programs are not only limited to biochemical testing, but also responsible for coordination of follow-up activities, including clinical care, education, and quality assurance. For a program to be successful, it must constantly evaluate its protocols and procedures to ensure its effectiveness. This involves engagement of various stakeholders, including government and nongovernment agencies, hospitals, academic centers, and private partners.

Metabolic testing

Laboratory evaluation of the IEMs includes both routine and specialized testing. Routine testing includes measurement of glucose, electrolytes, ammonia, lactate, pyruvate, renal function tests, liver function tests, urinalysis, and basic hematology tests. Most pediatric centers that offer specialized metabolic testing include amino acid, acylcarnitine, and urine organic acid analysis on their test menu to cover the most commonly detected IEMs on a state's newborn screening panel. The purpose of offering such services in-house is often to provide confirmatory testing on screen-positive infants, expanded coverage of disorders not included on the newborn screening panel, and faster results in cases where prompt therapeutic intervention is necessary. Some patients can be missed by newborn screening, especially for disorders that present later in life. Therefore a symptomatic patient at any age, regardless of newborn screening results, should be investigated for a potential IEM.

Specialized testing typically includes a combination of chromatography and mass spectrometry for the analysis of amino acids, carnitines (including free and total) and acylcarnitines, and organic acids. A combination of these tests is often needed to support a biochemical diagnosis of an IEM. Liquid chromatography−tandem mass spectrometry (LC−MS/MS) is commonly used for targeted analysis of small molecules including amino acids, acylcarnitines, and organic acids. Typically, compounds are separated by hydrophobicity with liquid chromatography and then filtered by unique mass-to-charge (*m/z*) ratios in mass spectrometers in tandem. Please refer to Chapter 10, Mass spectrometry, for a more detailed discussion on mass spectrometry. LC−MS/MS analysis is suitable for a wide variety of specimen types, including plasma, urine, cerebral spinal fluid (CSF), dried blood spots, and postmortem specimens. Gas chromatography−mass spectrometry (GC−MS) offers a broader coverage of metabolites.

Amino acid analysis

Amino acid analysis is helpful in evaluation of hyperammonemia, developmental delay, seizures, and metabolic acidosis. Specialized HPLC systems have been used to measure amino acids since the 1950s. Detection of amino acids by absorption or fluorescence requires deproteinization and derivatization that may be performed before or after separation. Fig. 48.1A illustrates an amino acid chromatogram generated by an ion-exchange chromatography method. Amino acids are separated based on their pK_a with a buffer gradient of increasing ionic strength. With this method, acidic amino acids are eluted first, followed by neutral and then basic amino acids. Ninhydrin is the most common postcolumn derivatization reagent for amino acids. It reacts with primary amino groups to produce a purple compound with an absorption maximum at 570 nm. The peak area for each compound is compared with an external calibration standard. Additionally, an internal standard is added to both the patient sample and calibrator to adjust for any variation in injection volume or reagent reactivity.

Amino acid analysis detects all amino-containing compounds, including nonamino acid compounds such as urea and ammonia. A number of other amino acids that are not found in proteins, but that aid in diagnosis of IEMs, are also identified. These include alloisoleucine, ornithine, citrulline, and argininosuccinic acid (ASA). Dietary artifacts and medications, as well as abnormal amino acids characteristic of various IEMs, may also be identified using this method. Identification of eluting compounds is based on retention time and relative elution order. Amino acid analysis with visible or fluorescent detection does not provide absolute identification of any compound. This is generally adequate for samples containing only normal physiologic amino acids, but the presence of medications or unusual amino acids due to an IEM complicate interpretation. Ion-exchange chromatography is the most commonly used method for amino acid analysis; however,

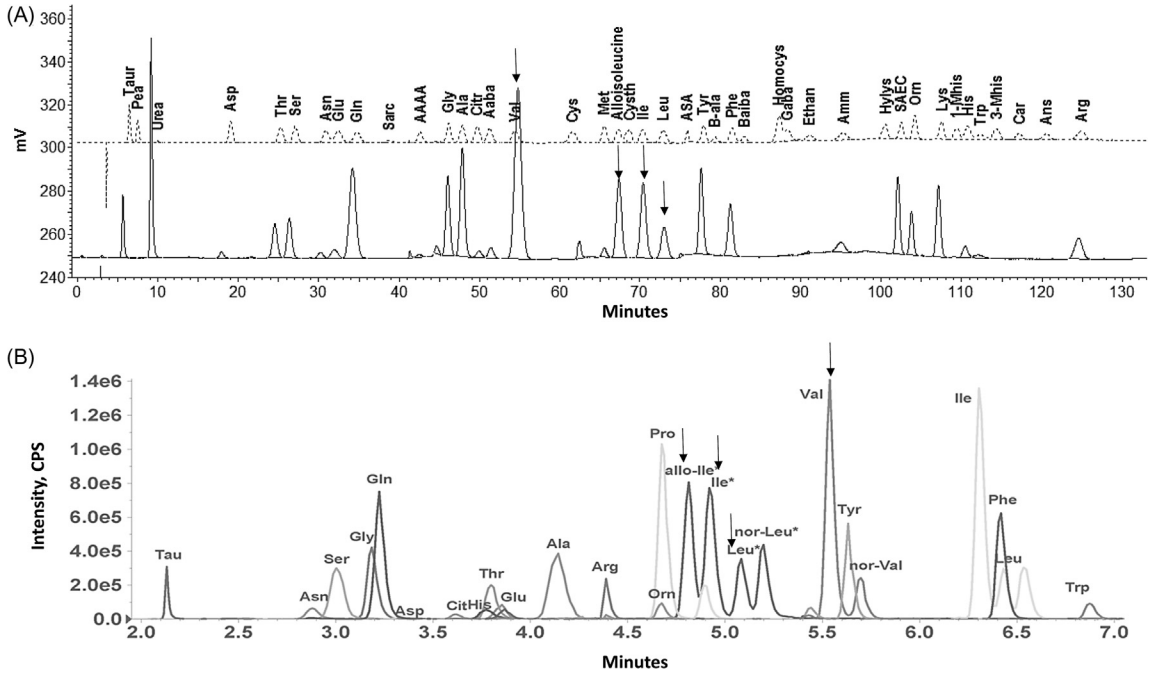


FIGURE 48.1 (A) Plasma amino acid analysis of a patient with maple syrup urine disease using an ion-exchange chromatography method. *Dotted line*: calibration standard. *Solid line*: patient sample. (B) Multiple reaction monitoring liquid chromatography−tandem mass spectrometry method. *Arrows*: diagnostic amino acids (alloisoleucine, isoleucine, leucine, and valine).

many laboratories are now starting to adopt LC−MS/MS methods [1,2]. LC−MS/MS assays for amino acid analysis offer more analytical specificity and sensitivity, as well as faster analysis time (Fig. 48.1B).

Plasma is the preferred sample for amino acid analysis for the diagnosis of most disorders. When possible, the blood should be collected in the fasting state (at least 3−4 hours postprandial) or after holding parenteral nutrition for 30−60 minutes. Plasma should be separated as soon as possible and frozen if analysis is delayed. If serum is used, the sample should be immediately centrifuged after clotting [3]. Cysteine and homocysteine attach to proteins in plasma samples and can be falsely decreased with improper storage conditions. Hemolysis will result in falsely increased concentrations of aspartate, glutamate, and taurine.

Urine amino acid analysis is usually less informative than analysis of plasma samples due to the wide reference intervals for urinary excretions of most amino acids. Reference intervals for urine amino acids are usually normalized for creatinine. Excretion of amino acids is especially high during infancy and decreases with age. In general, urine amino acid analysis should be reserved for assessment of renal tubular transport abnormalities and disorders of amino acid transport, such as lysinuric protein intolerance (characterized by an increase of lysine and other dibasic amino acids in urine), Hartnup disease, and cystinuria. Generalized dysfunction of proximal tubular reuptake (Fanconi syndrome), which results in elevations of many urine amino acids, may be seen in a number of IEMs, including mitochondrial disorders. CSF amino acid analysis is essential in the diagnosis of neurotransmitter disorders such as NKH, serine deficiency disorders, and pyridoxal-phosphate-dependent epilepsy.

Age-related reference intervals have been published for various sample types. Table 48.3 lists IEMs typically diagnosed by an amino acid profile and their associated amino acid abnormalities. Generally, glutamine, alanine, and valine are the largest peaks in the chromatogram. In urine, ammonia and urea predominate, whereas glycine and histidine are the amino acids usually present in the highest concentrations. Glutamine is the most abundant amino acid in normal, acellular CSF. Analysis of amino acids should be performed quantitatively, as qualitative analysis is not sufficient to identify some of the subtler changes in amino acid concentrations that characterize some disorders. Once an IEM is diagnosed, plasma amino acid analysis may be an important part of management, especially for urea cycle disorders (UCDs) and those conditions requiring restriction of selected amino acids.

Acylcarnitine analysis

Acylcarnitine analysis is helpful in the evaluation of hypoglycemia, hypotonia, cardiomyopathy, rhabdomyolysis, and metabolic acidosis. Analysis is commonly performed by MS/MS directly after derivatization to either butyl or methyl esters. Acylcarnitines arise from the conjugations of acyl-coenzyme A (CoAs) with carnitine for the transport of long-chain fatty acids across the innermitochondrial membrane for β-oxidation. Additionally, carnitine binds acyl residues and aids in their excretion. This mechanism is also important in the removal of organic acids that accumulate in some organic acidemias. For this reason, a plasma acylcarnitine profile can aid in the diagnosis of organic acidemas in addition to fatty acid oxidation disorders (FAODs). In certain IEMs, carnitine supplementation is employed to aid in the elimination of the toxic metabolite. Table 48.4 lists the acylcarnitine abnormalities that characterize various organic acid and FAODs.

Plasma is the preferred sample type for acylcarnitine analysis, although analysis of urine acylcarnitines can be useful in the workup of glutaric academia type I (GA1) when other tests have led to equivocal results. Hydrolysis of acylcarnitines in unfrozen samples can result in falsely decreased concentrations; therefore plasma samples should be processed immediately and kept frozen

TABLE 48.3 Inborn errors of metabolism associated with amino acid abnormalities.

IEM	Abnormal amino acid[a]
Maple syrup urine disease	↑Alloisoleucine, isoleucine, leucine, and valine
Lysinuric protein intolerance (low in serum)	Arginine, lysine, and ornithine (↑urine/↓serum)
Urea cycle defects	↑Glutamine (see Table 48.6)
Glycine encephalopathy	↑Glycine (serum and CSF)
Homocystinuria	↑Methionine and ↑homocysteine
Phenylketonuria	↑Phenylalanine and ↓tyrosine
Tyrosinemia types 1−3	↑Tyrosine
Hartnup disease	↑Neutral amino acids (urine)

IEM, Inborn errors of metabolism; *CSF*, cerebral spinal fluid.
[a]*Abnormality of amino acid found in serum unless indicated.*

TABLE 48.4 Acylcarnitines characteristic of fatty acid oxidation disorders.

Enzyme deficiency	Abnormal acylcarnitines (increased unless indicated)
Carnitine palmitoyltransferase I	Increased free carnitine (C0) and decreased C16 and C18 species
Carnitine transporter	Decreased C0, decreases in most acylcarnitines
Propionic acidemia	C3
Methylmalonic aciduria	C3, usually C4DC
SCAD	C4
SCHAD	C4-OH
Beta-ketothiolase	C5:1
Isovaleric acidemia	C5
Methylbutyrylglycinuria	C5
Glutaric acidemia type 1	C5DC
3-Methylglutaconic aciduria type I	C5OH and C6DC
MCAD	C6, C8, C10:1
VLCAD	C14:1, C14:0, C14:2, C16:0, C16:1, C18:1, C18:2, and C18:0
LCHAD	C16-OH, C16:1-OH, C18-OH, and C18:1-OH
Carnitine-palmitoyltransferase II	C16:0, C16:1, C18:1, C18:2, and C18:0; often normal
Carnitine-acylcarnitine translocase	C16:0DC, C18:0DC
Multiple acyl-CoA dehydrogenase deficiency	C4, C5 and C6-C18 acylcarnitines with and without double bonds; very low C2-carnitine
Mitochondrial disease	Elevations of multiple acylcarnitines in a nonspecific pattern
Renal insufficiency	Elevations of multiple acylcarnitines in a nonspecific pattern, more normal with hemodialysis, and unchanged by peritoneal dialysis
Fasting	Increased C2, C4−OH, and possibly other acylcarnitines

Notes: Acylcarnitines are named according to the number of carbons, the number of double bonds, and the presence or absence of a hydroxyl group as part of the acyl group. For example, octadeca-di-enoylcarnitine (18 carbons and two double bonds) is noted as C18:2. DC indicates the acylcarnitine derivative of the dicarboxylic acid of the corresponding number of carbons.
CoA, Coenzyme A; *LCHAD*, long-chain hydroxyacyl-coenzyme A dehydrogenase; *MCAD*, medium-chain acyl-coenzyme A dehydrogenase; *SCAD*, short-chain acyl-coenzyme A dehydrogenase; *VLCAD*, very long-chain acyl-coenzyme A dehydrogenase; *SCHAD,* short-chain hydroxyacyl CoA dehydrogenase.

until analysis. Hemolyzed specimens can lead to false increases in long-chain acylcarnitines and should not be used for analysis (or, at a minimum, should be interpreted with caution). Results may be reported as either a qualitative interpretation or as quantitative results. Qualitative results require careful examination, because some diagnostic acylcarnitines are only minimally increased. In either case, interpretation should be included with recommendations for additional testing indicated for further evaluation of abnormal results. Analysis by MS/MS without chromatographic separation identifies compounds based on the molecular mass and fragmentation; therefore these methods sometimes cannot distinguish isobaric compounds with similar chemical structures. For example, elevation of the 5-carbon acylcarnitine can be indicative of either isovalerylcarnitine or 2-methylbutyrylcarnitine. Additional studies (such as urine organic acids) are necessary to distinguish between potentially life-threatening isovaleric academia and the more benign 2-methylbutyrylglycinuria.

Although plasma acylcarnitine analysis allows rapid diagnosis of many conditions previously diagnosed by urine organic acid analysis, acylcarnitine analysis does not identify all of the compounds identified in the broader screening of urine organic acids. In addition, plasma acylcarnitine concentrations vary with feeding and fasting, and the profile can be normal in some FAODs when the sample is collected postprandially. For this reason, the excretions of organic acids characteristic of an IEM are often a better indicator of recent metabolic control than are acylcarnitine concentrations. Specificity is increased when analyses of organic acids and acylcarnitines are combined.

Urine organic acid analysis

GC−MS is used primarily for detection of organic acids, derived from metabolism. Compounds are extracted from urine into an organic solvent, derivatized to volatile forms, and separated by volatility and solubility in the stationary phase of the gas chromatography column. Some laboratories pretreat urine samples with urease to remove urea, which can interfere with the detection of other compounds, prior to trimethylsilyl derivatization [4]. The derivatization process enhances the volatility of organic acids. The mass spectrometer produces a characteristic mass spectrum, which is then matched against a library of

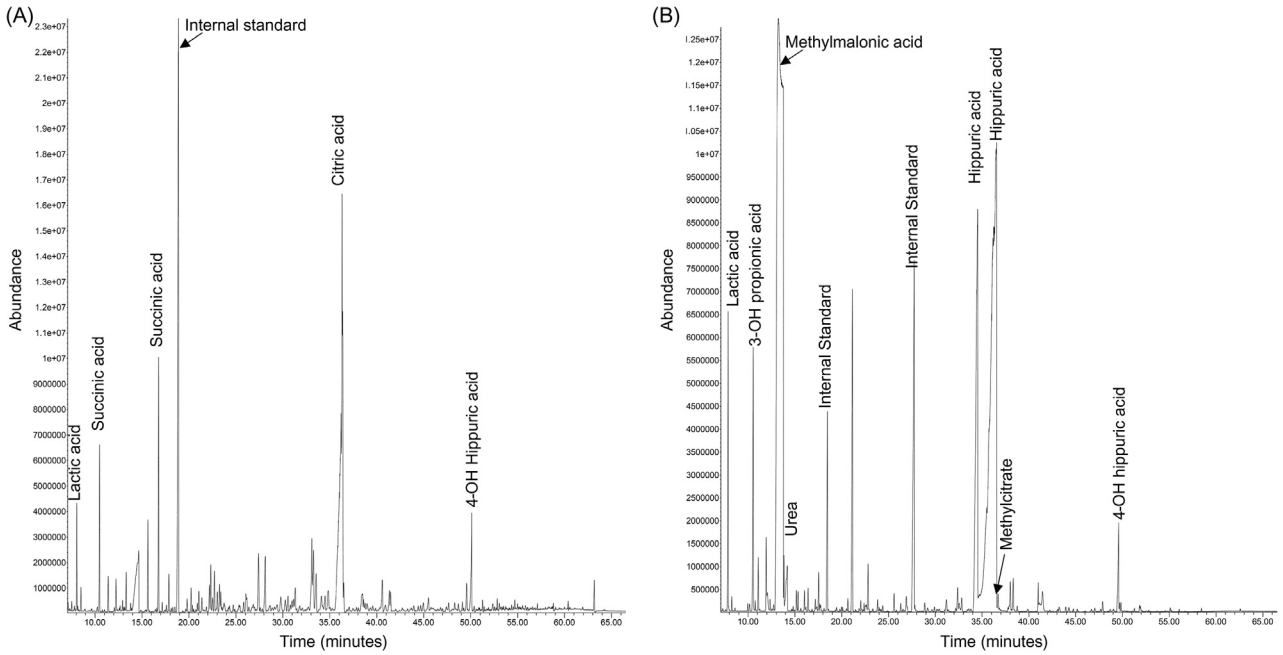


FIGURE 48.2 Urine organic acid profiles. Organic acids from urine after extraction with ethylacetate and trimethylsilyl derivatization. (A) Normal profile from a 2-week-old patient. (B) Urine from a patient with methylmalonic aciduria. Peaks are interpreted relative to the internal standard (2-phenylbutyric acid).

spectral data. Compounds are identified by both their *m/z* ratio as well as their retention time. Fig. 48.2 shows the urine organic acid profile of a healthy individual and a patient with MMA for comparison.

Urine is the preferred sample for screening organic acid analysis, because most abnormal metabolites are present in higher concentrations in urine than in plasma. Excretion of organic acids decreases with age as the renal tubules mature. Extraction of an amount of urine corresponding to a standard amount of creatinine makes it possible to visually compare profiles from one patient to the next, as well as several profiles obtained at various times for the same patient. Although quantitation is possible, it is usually done by comparing ion intensity for a particular compound to that of one or more internal standards. This provides only a semiquantitative estimate of the concentration because of the variations in extraction efficiency for different classes of compounds.

Once a particular IEM has been diagnosed, routine monitoring may include urine organic acid analysis or, in some conditions, measurement of the accumulated metabolites in plasma. Sensitivity of the analysis is increased by monitoring only a few ions characteristic of the compound in question [selected ion monitoring (SIM)] instead of the broad range of masses used in screening organic acid analysis. SIM/stable isotope dilution (SID) can also be used to quantitate selected organic acids when monitoring therapy in a patient previously identified with an IEM. Other methods that use tandem mass spectrometry and high-resolution mass spectrometry [hybrid quadrupole time-of-flight (TOF)] are used in some clinical laboratories. In some instances, the latter method offers both targeted quantitation and untargeted metabolite screening with increased sensitivity.

Extraction of urine organic acids for analysis can convert 2-ketoacids to their 2-hydroxy forms, thus impairing the detection of 2-ketoacids such as pyruvic acid and branched-chain 2-ketoacids found in maple syrup urine disease (MSUD), as well as, succinylacetone, the compound pathognomonic for tyrosinemia type 1. An oximation step with hydroxylamine hydrocline can be added to the procedure to help stabilize the keto-groups of these metabolites. Due to differences in methodology and instrumentation, a urine organic acids report from one laboratory may not exclude the same conditions excluded by another laboratory using different techniques. In addition, the organic acids characteristic of some conditions may not be present in urine when the patient is asymptomatic, and it may be necessary to repeat the analysis. In some disorders, the characteristic organic acids may be present in very low concentrations, and accurate quantitation is necessary.

The identification of compounds also depends on the quality of the reference library used for spectral data. It is recommended that laboratories periodically review or update these libraries to ensure that key diagnostic

metabolites are not missed in the screening procedure. The interpretation of organic acid profiles can be challenging due to the number of compounds that are detected in urine samples as well as the presence of interferences and artifacts produced from diet and medications. Ideally, the interpretation unites the results of urine organic acid analysis with the rest of the patient's clinical history.

Other metabolic testing

Enzyme activity or genetic testing can be used in the setting of equivocal biochemical findings for confirmation of a diagnosis. Specimens for enzyme activity testing may include skin fibroblasts or lymphocytes. In some cases, enzyme testing may involve invasive procedures to obtain the cell type with the enzyme defect and should be used as a last resort for diagnosis. DNA testing by either targeted mutation analysis or whole gene sequencing can be beneficial. Genotype−phenotype correlations can help in directing therapy and predicting disease course. As the genetic basis of many IEMs has been determined, genetic testing is being used increasingly to confirm diagnosis and perform carrier testing. One limitation of gene sequencing is that it may not identify all disease-causing mutations, and it can miss single-exon deletions and duplications. Additionally, in some cases, identification of variants of unknown significance can cause further ambiguity in cases where the clinical and biochemical diagnoses are uncertain. Of note, some newborn screening programs have now incorporated second-tier molecular testing to identify common disease-causing variants for faster diagnosis and confirmation of certain disorders such as galactosemia, medium-chain acyl-coenzyme A dehydrogenase (MCAD) deficiency, and very long-chain acyl-coenzyme A dehydrogenase (VLCAD) deficiency.

The disorders

Amino acid disorders

Disorders of amino acid metabolism are generally disorders of intoxication, due to dysfunctional catabolism of the carbon skeleton or dysfunctional transport across cellular bodies. The clinical presentation is highly variable and nonspecific, and the laboratory diagnosis is based on the accumulated metabolite. In most instances, plasma amino acid and urine organic acid analyses aid in the diagnosis of these disorders.

Phenylketonuria

PKU was the first disorder evaluated by newborn screening. Classical PKU is due to a complete deficiency in phenylalanine hydroxylase (PAH), which converts phenylalanine to tyrosine (Fig. 48.3). In PKU, increased phenylalanine and decreased tyrosine lead to neurotoxicity as well as impaired neurotransmitter production. Untreated PKU leads to severe mental retardation that can be prevented by adherence to a phenylalanine-restricted diet.

PKU is typically diagnosed by elevated plasma phenylalanine concentrations with a low or low-normal tyrosine. Additionally, excretion of phenyllactate, phenylpyruvate, and phenylacetate can be identified by urine organic acid analysis in patients with PKU; however, it is important to note that these metabolites may be commonly observed in newborns with immature livers. The majority of hyperphenylalaninemias are caused by mutations in the *PAH* gene, in which more than 400 variants have been identified. In the remainder (2% of cases), hyperphenylalaninemia is secondary to a deficiency in tetrahydrobiopterin (BH4), a cofactor required by PAH. Measurement of urinary pterins and blood dihydropteridine reductase activity is recommended to exclude defects in BH4 before

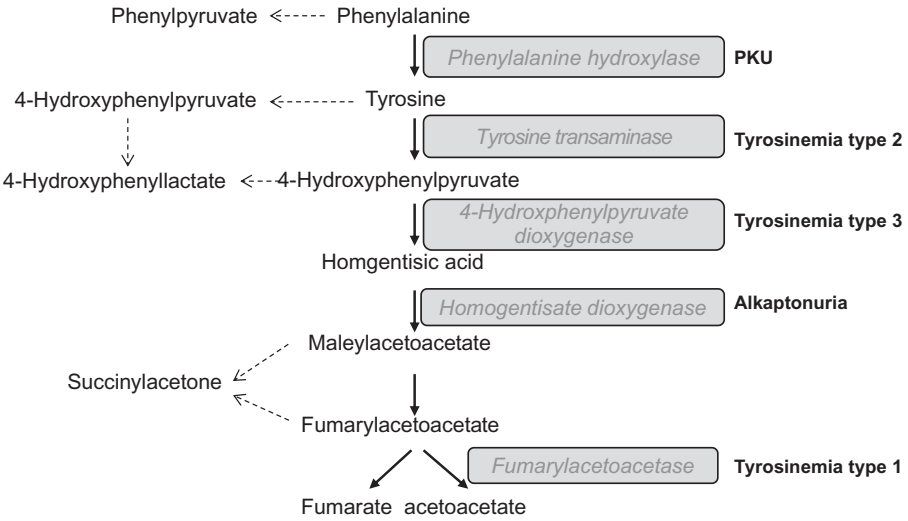


FIGURE 48.3 Disorders of phenylalanine and tyrosine metabolism. *Bold*: inborn errors of metabolism resulting from deficiencies in enzymes. *Dashed arrows*: abnormal production of metabolites. *PKU*, Phenylketonuria.

initiation of treatment as BH4 deficiency, as treatment course will vary depending on the type of hyperphenylalaninemia. Approximately 35% of all hyperphenylalaninemias are responsive to BH4 supplementation and allow for increased phenylalanine tolerance in the diet. Depending on the type of BH4 deficiency, patients may also require neurotransmitter replacement therapy. Treatment of PKU with a phenylalanine-restricted diet requires frequent monitoring with serial plasma amino acid analysis to ensure that the phenylalanine concentration is low enough to prevent brain damage, but not so low that phenylalanine deficiency results in growth retardation. Elevated maternal phenylalanine concentrations during pregnancy can result in serious complications, including intrauterine growth restriction, microcephaly, intellectual disability, and even fetal loss [5].

Hereditary tyrosinemias

Elevation of plasma tyrosine is most often due to liver disease, but it is also found in varying degrees in the inherited abnormalities of tyrosine metabolism (tyrosinemia types 1, 2, and 3, as well as alkaptonuria and hawkinsinuria). Furthermore, some newborns have transient tyrosinemia due to delayed maturation of 4-hydroxyphenylpyruvate dioxygenase. There are a number of steps in the metabolism of tyrosine. Deficiencies in the metabolic steps closest to tyrosine result in the highest elevation of plasma tyrosine. 4-Hydroxyphenyllactic acid and 4-hydroxyphenylpyruvic acid can be detected by urine organic acid analysis in all of the hereditary tyrosinemias, and it is not possible to distinguish among the hereditary tyrosinemias or between tyrosinemia due to liver disease or immaturity and hereditary tyrosinemia based on these compounds (Fig. 48.3). The presence of succinylacetone is pathognomonic for tyrosinemia type 1; however, its excretion may be so low that it cannot be identified except using SIM/SID. It is important to identify individuals with tyrosinemia type 1, because treatment with a tyrosine- and phenylalanine-restricted diet can reverse their liver disease, but most will develop hepatocellular carcinoma because of the carcinogenic activity of fumarylacetoacetate. Treatment with nitisinone (NTBC), an inhibitor of 4-hydroxyphenylpyruvate dioxygenase, prevents the accumulation of fumarylacetoacetate and dramatically reduces the risk of hepatocellular carcinoma.

Maple syrup urine disease

Deficiency of branched-chain ketoacid dehydrogenase (BCKDH) results in the accumulation of branched-chain 2-ketoacids and branched-chain amino acids (Fig. 48.4). MSUD is biochemically characterized by elevated blood concentrations of isoleucine, leucine, and valine, as well as increased urinary excretion of corresponding α-ketoacids: α-keto-3-methylvalerate, α-ketoisocaproate, and α-ketoisovalerate, which can be identified by urine organic acid analysis. Most importantly, the detection of alloisoleucine on plasma amino acid analysis is virtually pathognomonic for MSUD. Although newborn screening is available for MSUD, infants with classic MSUD are often already ill when the results of the newborn screening test become available. The marked elevation of leucine is responsible for the lethargy and coma that are the typical presentation of MSUD.

In classical MSUD, routine laboratory testing at presentation is often unremarkable, and urine dipsticks for ketones are usually negative in newborns. Urine does not usually have the characteristic maple syrup odor in newborns, but the odor may be present in cerumen. Amino acid analysis shows marked elevation of branched-chain amino acids; a typical profile is shown in Fig. 48.1. Although branched-chain amino acids can be markedly increased with prolonged fasting or episodes of vomiting, MSUD can be confirmed by the presence of alloisoleucine. It is important to note that medications with amines in their chemical structures, such as antiepileptics, can

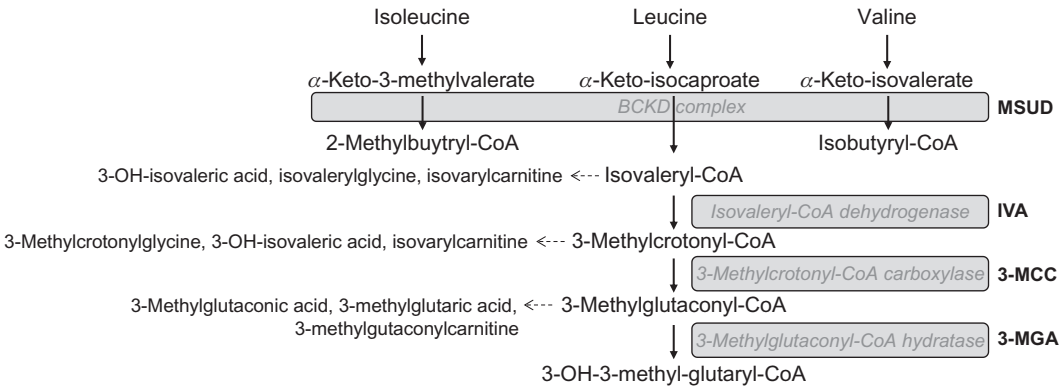


FIGURE 48.4 Disorders of branched-chain amino acid metabolism. *Bold*: inborn errors of metabolism resulting from deficiencies in enzymes. *Dashed arrows*: abnormal production of metabolites. *3-MCC*, 3-Methylcrotonyl-coenzyme A carboxylase; *IVA*, isovaleric aciduria; MGA: 3-methylglutaconic aciduria; *MSUD*, maple syrup urine disease.

coelute with alloisoleucine, resulting in falsely increased concentrations when an MS/MS-based detection method is not employed. Severe deficiency of BCKDH results in classical MSUD. Intermediate MSUD may not be diagnosed until later in life due to residual BCKDH activity, which can range from 3% to 40% in affected individuals. Some individuals with MSUD have enough residual activity of BCKDH that they have only intermittent symptoms with normal plasma amino acid concentrations between episodes. However, even individuals with intermediate or intermittent, MSUD can experience severe neurotoxicity if there is sufficient metabolic stress to overload the residual BCKDH activity.

Treatment of MSUD requires a diet restricted in the branched-chain amino acids: valine, isoleucine, and leucine. It is often necessary to supplement valine and isoleucine to maintain leucine at the desired concentration. As with other IEMs requiring a protein-restricted diet, regular monitoring is necessary to prevent amino acid deficiencies while maintaining branched-chain amino acids low enough to prevent toxicity.

Homocystinuria

The sulfur-containing amino acid cysteine is synthesized from methionine. This pathway proceeds via homocysteine, which can be remethylated to methionine. Homocystinuria can result from a number of enzyme deficiencies in this pathway, which requires folic acid, vitamin B_{12}, and pyridoxine (vitamin B_6). Even a small amount of homocysteine detected by plasma amino acid analysis is abnormal and indicative of disease. Various forms of homocystinuria can be distinguished by plasma methionine and the presence or absence of methylmalonic aciduria (MMA).

Classical homocystinuria is due to a deficiency in cystathionine-β-synthase (CBS) activity, and results in marked elevation of methionine in addition to homocystine; cysteine concentration may be decreased. Homocysteine is also increased in several disorders of folic acid and vitamin B_{12} metabolism, such as cobalamin C disease. In these conditions, methionine concentration is low. The combination of homocystinuria and MMA is found in two disorders of vitamin B_{12} metabolism.

Screening for the increased blood methionine seen in homocystinuria due to CBS deficiency is included in the newborn screening panel of many states. A number of other conditions can result in increased plasma methionine including liver disease, increased protein intake, and methionine adenosyltransferase deficiency. Newborn screening may not identify individuals with milder forms of CBS deficiency, because the blood methionine concentration is often not increased sufficiently in the first few days of life. Instead, these individuals often present with lens dislocation or thromboembolic event later in life. The increased risk of clotting episodes in homocystinuria led researchers to investigate milder elevations of homocysteine as a risk factor for strokes and myocardial infarction.

Diagnosis and management of all forms of homocystinuria require measurement of homocysteine in addition to plasma amino acids. The sulfhydryl groups of cysteine and homocysteine readily react with the free sulfhydryl groups of plasma proteins. Therefore most of the circulating homocysteine is found in homocysteine-cysteine or homocysteine dimers (10%−20%) or is protein bound (80%), and only 1%−2% is free or unbound homocystine. If amino acid analysis of a plasma sample is delayed, free cystine (cysteine dimers) and homocystine (homocysteine dimers) decrease, and the diagnosis may be missed. Instead, measurement of total plasma homocysteine provides a more reliable indication of homocystinuria. Immunoassays for total homocysteine are available on most automated chemistry/immunoassay systems. The general assay principle involves reducing the circulating forms of homocysteine to free homocysteine using reducing agents. Free homocysteine is then converted to S-adenosylhomocysteine (SAH), which serves as the analyte for a competitive immunoassay. These assays are more commonly used for assessment of cardiovascular risk, which is associated with mild hyperhomocystinemia. Please refer to Chapter 29, Pediatric laboratory medicine, for a discussion of laboratory testing related to cardiovascular disease. The analytical range of some automated tests for total homocysteine may not be adequate to allow accurate quantitation of the markedly increased total homocysteine found in individuals with CBS deficiency, and it may be necessary to dilute such samples.

Treatment of severe CBS deficiency centers around reduction of plasma homocysteine concentrations by a methionine-restricted diet, high-dose vitamin B_6, or betaine administration. Betaine lowers plasma homocysteine concentrations by donating a methyl group to homocysteine to generate methionine. Treatment decreases the risk of thrombosis as well as the mild developmental disabilities that can be seen in untreated, severe CBS deficiency. In contrast, homocystinuria due to disorders of folate and vitamin B_{12} is frequently associated with megaloblastic anemia, microcephaly, and more severe developmental disabilities, possibly due to inadequate supply of *S*-adenosylmethionine. In these conditions, treatment with betaine can decrease total homocysteine and increase plasma methionine.

Other amino acid disorders

A number of amino acid disorders can result in severe neonatal seizures. These include NKH, serine deficiency disorders, and pyridoxal-phosphate-dependent epilepsy.

CSF amino acid analysis is usually necessary to diagnose or exclude these conditions, because the results of plasma amino acid analysis may be normal or nonspecific. In NKH, plasma glycine is usually markedly increased, but this finding is not specific. Instead, diagnosis requires an elevation of CSF glycine and an increased CSF glycine/plasma glycine ratio (reference <0.08). Infants with serine deficiency disorders may have normal plasma serine, whereas CSF serine is uniformly decreased (6−8 μM; normal range 35−80 μM). Pyridoxal-phosphate-sensitive encephalopathy results in elevation of CSF threonine with mild elevation of glycine.

Sulfite oxidase deficiency is another cause of neonatal seizures. It can occur as an isolated deficiency or combined with xanthine oxidase deficiency due to molybdenum cofactor deficiency. Plasma taurine is increased, but this finding is common in formula-fed infants. Sulfocysteine can also be increased with decreased plasma cystine, but sulfocysteine is unstable and not included with amino acid standards and may be missed without careful inspection of the amino acid profile. Low plasma cystine is almost always due to delay in amino acid analysis and is not specific enough to be indicative of the need for additional testing. Diagnosis of sulfite oxidase deficiency is most often suggested by increased urine sulfite and extremely low plasma total homocysteine.

Organic acid disorders

Organic acidemias are a diverse group of IEMs, in which one or more nonamino acids accumulate in tissues and body fluids. In most organic acidemias, the characteristic organic acid(s) are present in urine; thus urine organic acid analysis by GC−MS is currently the gold standard for the identification of these disorders. Organic acids are derived from amino acid metabolism, ketone body and fatty acid metabolism, and Krebs cycle precursors and intermediates. In patients with metabolic defects in these pathways, both precursors and their side products or metabolites can be detected by urine organic acid analysis [6]. For instance, in patients with propionic academia, deficiency in the enzyme propionyl-CoA carboxylase results in the accumulation of propionyl-CoA, which further reacts with oxaloacetate to produce methylcitrate, a metabolite that is not present in the urine of healthy individuals.

The presence of a single organic acid at high concentrations is not necessarily diagnostic of an IEM. The excretion of these metabolites in the urine can vary based on protein intake and other factors. Ideally, the interpretation, essential for those other than metabolic specialists, unites the results of urine organic acid analysis with other laboratory findings as well as the patient's clinical history. If the sample is from a patient previously diagnosed with an IEM, it may be helpful to compare the profile of organic acids in the current sample with the results obtained in previous samples. With some organic acid disorders, specific plasma acylcarnitine patterns can aid in diagnosis as well as monitoring. Table 48.5 lists common organic acidurias as well as their associated urine metabolites and clinical features.

The most classic organic acid (OA) disorders result from defects in amino acid degradation. These disorders typically manifest in the newborn period, but can also present later in life. Vomiting, poor feeding, and lethargy are typical manifestations in the neonatal period. Milder forms of OA disorders are associated with repeat episodes of metabolic acidosis, hypoglycemia, and leukopenia. Some types of OA disorders have severe neurologic manifestations and present with acute encephalopathy, seizures, and ataxia. Below, we provide a few examples of commonly seen OA disorders.

3-Methylcrotonylglycinuria

3-Methylcrotonylglycinuria (3MCC) is the most frequent OA disorder detected by newborn screening in the United States, Europe, and Australia. It results from a deficiency in a biotin-dependent carboxylase, 3-α-methylcrotonyl-CoA carboxylase, involved in the degradation of leucine (Fig. 48.4). The presentation can range from severe neurological impairment to completely asymptomatic patients who only have a biochemical abnormality. The diagnosis is based on increased urinary excretion of 3-hydroxyisovaleric acid and 3-methylcrotonylglycine in the urine and increased plasma 3-hydroxyisovalerylcarnitine (C5-OH).

Isovaleric aciduria

Deficiency of isovaleryl-CoA dehydrogenase, an enzyme involved in the leucine degradation pathway, results in isovaleric aciduria (IVA; Fig. 48.4). A characteristic odor resembling "sweaty feet" caused by the accumulation of isovaleric acid has been described in patients. Typical manifestations include acidosis, ketosis, vomiting, lethargy, and coma. The diagnostic feature of IVA is elevated concentrations of isovalerylglycine and 3-hydroxyisovaleric acid in the urine; however, the latter is found only during metabolic decompensation. Elevations in plasma isovaleryl carnitine (C5) can also aid in the diagnosis and monitoring, and is used as the marker for newborn screening. Treatment involves a low protein, leucine-reduced diet with carnitine, and glycine supplementation.

Propionic acidemia and methylmalonic aciduria

Propionic acidemia (PA) and MMA are organic acid disorders due to defects in catabolism of isoleucine, valine,

TABLE 48.5 Urine metabolites and clinical presentation of selected organic acidurias.

Organic aciduria	Urine metabolites	Clinical presentation
Propionic acidemia	3-Hydroxypropionic acid Propionylglycine Methylcitrate Lactic acid	Poor feeding, vomiting, ketoacidosis, lethargy, coma, seizures, and cardiomyopathy
Methylmalonic aciduria	Methylmalonic acid 3-Hydroxypropionic acid Propionylglycine Methylcitrate Lactic acid	Poor feeding, vomiting, ketoacidosis, lethargy, coma, seizures, cardiomyopathy, and renal failure
Biotinidase deficiency	3-Hydroxypropionic acid Propionylglycine Methylcitrate Lactic acid 3-Methylcrotonylglycine 3-Hydroxoisovaleric acid	Seizures, ataxia, progressive psychomotor retardation, skin rash, and alopecia
Isovaleric acidemia	Isovalerylglycine 3-Hydroxyisovaleric acid	Poor feeding, vomiting, "sweaty feet" odor, lethargy, coma, cerebral edema, and ketoacidosis
3-Methylcrotonyl-CoA carboxylase deficiency	3-Methylcrotonylglycine 3-Hydroxyisovaleriac acid	Poor feeding, vomiting, ketoacidosis, lethargy, coma, and psychomotor retardation
Glutaric aciduria type I	Glutaric acid 3-Hydroxyglutaric acid Glutaconic acid	Macrocephaly, acute encephalopathy, hypotonia, and impaired neurodevelopment
3-Methylglutaconic aciduria type I	3-Methylglutaconic acid 3-Methylglutaric acid 3-Hydroxyisovaleric acid	Fasting hypoglycemia, delayed speech, and progressive psychomotor retardation

CoA, Coenzyme A.

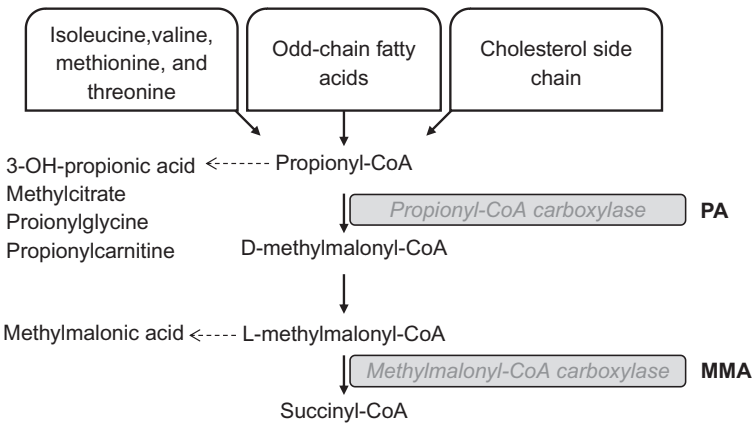


FIGURE 48.5 Metabolism of propionyl-coenzyme A and methylmalonyl-coenzyme A. Bold: inborn errors of metabolism resulting from deficiencies in enzymes. Dashed arrows: abnormal production of metabolites. *MMA*, Methylmalonic aciduria; *PA*: propionic acidura.

methionine and threonine, odd chain fatty acids, and cholesterol side chains (Fig. 48.5). Since this catabolic pathway provides succinyl-CoA to the Krebs cycle, PA and MMA are disorders of energy metabolism. Patients typically present with hypotonia, lethargy, and poor feeding. Decompensation can be rapid and severe. PA is caused by deficiency in propionyl-CoA carboxylase (whether primary or as part of multiple carboxylase deficiency). PA is diagnosed by presence of 3-hydroxypropionic acid, methylcitrate, proprionylglycine, and tiglylglycine in urine organic acid analysis. MMA refers to a group of disorders that result in the accumulation of methylmalonic and propionic acids caused by either methylmalonyl-CoA mutase deficiency, 5′-deoxyadenosylcobalamin metabolic defects, or vitamin B_{12} deficiency. Urine organic acid analysis of the MMA patient reveals elevations of methylmalonic acid, methylcitrate, and 3-hydroxypropionic acid. In both PA and MMA, elevation of C3 carnitine will be present in acylcarnitine analysis and is the marker used for newborn screening. PA and MMA are treated with protein-restricted diets fortified with amino acid supplements that are low in PA precursors, and carnitine supplementation.

Glutaric academia type I

Defects in glutaryl CoA dehydrogenase, an enzyme involved in the degradation of lysine, hydroxylysine, and tryptophan, results in GA1. Patients typically present with acute life-threatening encephalopathic crises usually within the first year of life. Diagnosis is based on detection of increased concentrations of glutaric acid, 3-hydroxyglutaric acid (characteristic of GA1), and glutaconic acids in the urine. Elevated glutarylcarnitine in plasma is also found and is used as the basis of newborn screening. Prompt recognition of this condition and restriction of lysine and tryptophan with carnitine supplementation can prevent neurological damage.

Urea cycle disorders

The urea cycle is important for the elimination of nitrogen from protein metabolism. Deficiency of each of the urea cycle enzymes has been described. In the urea cycle, ammonia is converted to urea through a series of reactions occurring in the mitochondria and cytosol (Fig. 48.6). Deficiencies of mitochondrial enzymes are collectively termed "proximal" UCDs, whereas the distal disorders involve cytosolic enzymes. Typically, affected individuals present with lethargy and coma within the first few days of life, but usually after the first 24 hours. Milder deficiencies may not present until much later in life. Males with X-linked ornithine transcarbamylase (OTC) deficiency may not become symptomatic until well into adult life, when their metabolic status is stressed by increased protein intake or increased protein turnover secondary to trauma, infection, or chemotherapy. Female carriers of OTC mutations that result in severe deficiency may also become symptomatic during times of increased metabolic stress. A common presentation is unexplained postpartum coma.

Hyperammonemia is a common presenting biochemical abnormality in individuals with UCDs. Individuals with UCDs often have very low blood urea nitrogen (as low as 0−1 mg/dL), but this finding is not specific for UCDs. Some UCDs can be diagnosed by amino acid analysis alone. Marked elevation of citrulline is indicative of argininosuccinate synthase deficiency. Milder elevation of citrulline is seen in argininosuccinate lyase deficiency where ASA also accumulates. Amino acid analyzers are often not programmed to identify ASA and its anhydrides, which may erroneously be identified by the instrument as other amino acids such as ethanolamine. Arginase deficiency is unique among the UCDs in that it is associated with only mild increases in plasma ammonia concentration. Affected patients often do not display symptoms of hyperammonemia and typically present outside the neonatal period with spastic diplegia, seizures, and cognitive deficits. Marked elevation of arginine is a hallmark of arginase deficiency, but it can also be seen in patients with other UCDs who are treated with IV arginine to replenish urea cycle intermediates during an episode of metabolic decompensation.

Increased plasma glutamine is a common finding in hyperammonemia, including postprandial hyperammonemia, and glutamine can be massively increased in UCDs. Elevation of plasma glutamine and alanine are common with deficiency of carbamyl phosphate synthase I (CPS I)

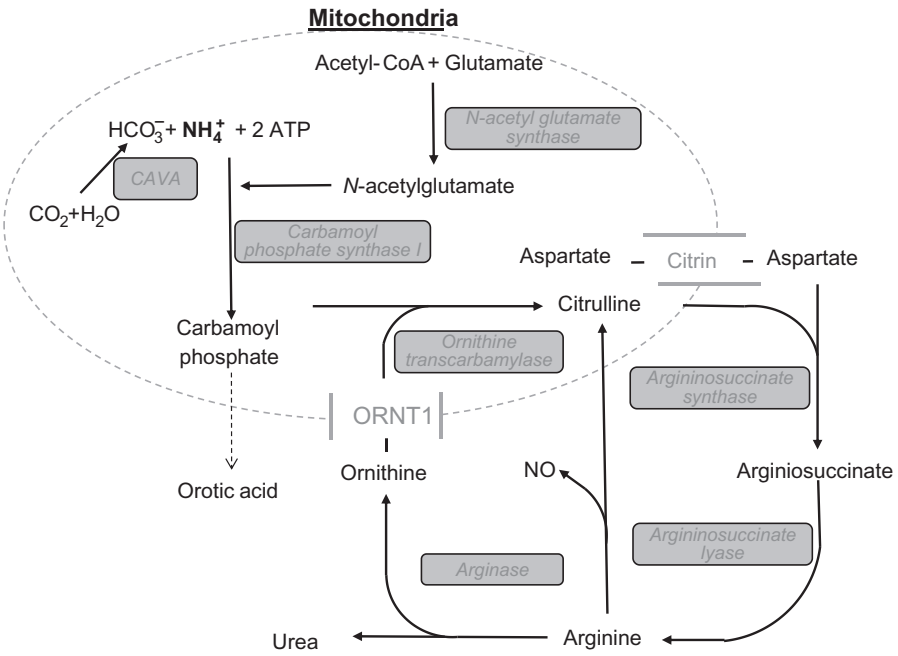


FIGURE 48.6 Urea cycle disorders. Urea cycle disorders involve the "proximal" reactions that occur in the mitochondria or "distal" reactions that occur in the cytoplasm as well as two transporters, citrin and mitochondrial 1 ornithine transporter. Carbonic anhydrase 5A produces the bicarbonate used by carbamoyl phosphate synthase I. In addition to its role in the urea cycle, arginine also serves as the source of nitric oxide.

and OTC. Plasma citrulline is often decreased, but ornithine and arginine are usually normal. Amino acid analysis cannot distinguish between CPS I and OTC deficiencies. Instead, organic acid analysis is needed to identify the increased orotic acid seen in OTC deficiency that results from accumulation of carbamyl phosphate in enzyme deficiencies after CPS I.

Massive elevation of ammonia may require dialysis initially. Once stabilized, ammonia-conjugating agents such as benzoic acid, phenylacetate, and phenylbutyrate allow additional diversion of ammonia from the urea cycle. Treatment of UCDs often includes protein restriction to decrease production of excess ammonia. Some dietary protein may be substituted by essential amino acid mixtures to allow higher total protein intake and improved growth. Supplementation with citrulline or arginine may also be necessary to replenish urea cycle intermediates. Despite optimal therapy, infants with neonatal hyperammonemia often have developmental delay when ammonia is markedly increased (>200 fmol/L) in the neonatal period. The enzyme deficiencies resulting in UCDs and their associated biochemical findings are listed in Table 48.6.

Fatty acid oxidation disorders

Beta-oxidation of most fatty acids occurs in the mitochondrion. Long-chain fatty acids are transported across the mitochondrial membranes by conjugation to carnitine. Once the acylcarnitines reach the mitochondrial matrix, the fatty acid is transferred back to CoA to form acyl-CoA thioesters. In FAODs, metabolism of acyl-CoAs is blocked, and acyl-CoAs accumulate with transfer of the fatty acid back to carnitine. The resulting accumulation of acylcarnitines can be analyzed by MS/MS. This testing can identify elevations of other CoA compounds, including acetyl-CoA that is increased with lipolysis and lactic acidosis and the short-chain acyl-CoAs that accumulate in organic acidemias. Fatty acid oxidation results in the production of acetyl-CoA that can be used to produce ketone bodies when glucose supply is inadequate. The enzymes involved in fatty acid oxidation are active on fatty acids of specific chain length. In FAODs, the chain length of the accumulating fatty acid intermediates influences the symptoms seen in these conditions.

Long-chain fatty acids are sequentially metabolized by VLCAD and long-chain hydroxyacyl-CoA dehydrogenase. Deficiencies of either of these enzymes may present with cardiomyopathy and myopathy due to energy deficiency in muscle, which preferentially uses fatty acids for fuel, or hepatic dysfunction due to accumulation of toxic acyl-CoA and acylcarnitine intermediates. Less severe deficiencies may present with hypoketotic hypoglycemia. Episodes of rhabdomyolysis are common after exercise or with metabolic stress and may be the presenting symptom in the mildest forms of these conditions. Long-chain FAODs are most often diagnosed by acylcarnitine analysis, because urine organic acid analysis may be normal and often shows only hypoketotic dicarboxylic aciduria when it is abnormal. This pattern, though suggestive of a FAOD, does not point to a specific diagnosis. Additionally, acylcarnitine analysis may be normal when

TABLE 48.6 Urea cycle disorders and associated biochemical findings.

	Disorder (defect)	Biochemical findings
Proximal UCD	*N*-acetyl glutamate synthase deficiency	Hyperammonemia, low citrulline, and arginine
	Carbamyl phosphate synthetase I deficiency	Hyperammonemia, low citrulline, and arginine
	Carbonic anhydrase 5A deficiency	Hyperammonemia, low citrulline, and arginine (lack of mitochondrial bicarbonate production also affects pyruvate carboxylase, propionyl-CoA carboxylase, and 3MCC-carboxylase)
	Ornithine transcarbamylase deficiency	Hyperammonemia, low citrulline and arginine elevated orotic acid, and uracil (urine)
Distal UCD	Argininosuccinate synthetase 1 deficiency or citrullinemia type I	Hyperammonemia, very high citrulline and low arginine, and elevated orotic acid (urine)
	Argininosuccinate lyase deficiency	Hyperammonemia, high citrulline and low arginine, ASA (plasma or urine), and elevated orotic acid (urine)
	Arginase deficiency	Mild hyperammonemia, elevated arginine, and elevated orotic acid (urine)
Transporter defects	ORNT1 deficiency or hyperornithinemia, hyperammonemia, and homocitrullinuria syndrome	Hyperammonemia, high ornithine, and high homocitrulline
	Citrin deficiency or citrullinemia type II	Mild-to-severe hyperammonemia and elevated citrulline

Notes: Because arginine supplementation is standard of care for many urea cycle disorders, arginine may be elevated due to therapeutic intervention.
ASA, Argininosuccinic acid; CoA, coenzyme A; *3MCC*, 3-methylcrotonylglycinuria; *ORNT1*, mitochondrial 1 ornithine transporter; *UCD*, urea cycle disorder.

the patient is asymptomatic, and acylcarnitine profile of cultured fibroblasts may be necessary for diagnosis.

Carnitine is required to transport long-chain fatty acids into the mitochondria. Fatty acids are transferred from CoA to carnitine by carnitine palmitoyl transferase I (CPTI) at the outer mitochondrial membrane. The acylcarnitines are transported across the mitochondrial intermembrane space by carnitine-acylcarnitine translocase (CACT), which counter transports free carnitine back to the outer mitochondrial membrane. Once the fatty acid enters the mitochondrial matrix, it is then transferred back to CoA by the action of carnitine palmitoyl transferase II (CPTII). Abnormalities in each of these enzymes as well as the active transport of carnitine have been described.

Because carnitine is required for metabolism of long-chain fatty acids, deficiencies of CPTI, CPTII, CACT, and the carnitine transporter typically have symptoms similar to those seen with long-chain fatty acid disorders. Severe deficiencies often present with cardiomyopathy or arrhythmia in the neonatal period, whereas individuals with milder deficiencies may have muscle weakness and episodic rhabdomyolysis. Plasma acylcarnitines may be diagnostic in more severe deficiencies of CPTI, CPTII, and CACT, but this testing may be unremarkable in mild deficiencies. An elevated free carnitine fraction, in the setting of low acylcarnitine and elevated (or high normal) total carnitine, may suggest a CPTI deficiency.

MCAD deficiency is the most common FAOD. It is often identified by abnormal newborn screening acylcarnitine profile, but it may present with hypoketotic hypoglycemia. Acylcarnitine analysis usually shows a characteristic profile (typically elevated C6, C8, C10:1), but it can be normal. Urine organic acid analysis can demonstrate the presence of hexanoylglycine (HG) and suberylglycine (SG), even when the patient is asymptomatic. These metabolites can also be quantitated by SID/SIM as part of acyl glycine analysis, and low excretion of HG and SG may require SID/SIM in some individuals with MCAD deficiency.

Short-chain acyl-coenzyme A dehydrogenase (SCAD) deficiency is difficult to diagnose, mostly because of continuing controversy about the clinical features of this condition. Some children have presented with neurologic problems and have shown elevations of ethylmalonic acid and methylsuccinic acid on organic acid analysis. However, most individuals with elevation of urinary ethylmalonic acid excretion do not have mutations in the SCAD gene. Some have a common polymorphism that results in mild reduction in SCAD activity. Research continues to determine the clinical significance of this polymorphism. In contrast to other FAODs, deficiency of short-chain 3-hydroxyacyl-CoA dehydrogenase results in hyperinsulinemic hypoglycemia.

Deficiency of multiple acyl-CoA dehydrogenase (MAD) enzymes (also known as Glutaric acidemia Type 2) is secondary to abnormalities of either electron transport flavoprotein (ETF) or ETF-coenzyme Q10 oxidoreductase (ETF-QO). ETF transfers electrons from the fatty acid dehydrogenases to the respiratory chain. ETF-QO is required to reoxidize ETF. Severe forms of MAD deficiency present in the neonatal period with dysmorphic facies and renal cysts and are often fatal in the neonatal period, but milder deficiency of MAD may be asymptomatic if fasting is avoided. Laboratory findings include fasting hypoketotic hypoglycemia, excretion of glutaric, ethylmalonic and dicarboxylic acids in the urine, and increased C4−C8 plasma carnitines.

Mitochondrial disorders

Mitochondria play an important role in production of ATP and found in large numbers in cells and tissues with high energy demand. Mutations in either mitochondrial DNA (mtDNA) or nuclear DNA can lead to mitochondrial dysfunction. Affected individuals with mutations in mtDNA can display mosaicism within a cell with a mixture of both wild-type and mutated mtDNA (heteroplasmy). The percentage of mutated mtDNA can vary among tissues and organs within the same individual as well as individuals within the same family. This variation may explain the varied phenotypes observed in patients with mtDNA disorders. Additionally, with more than 1000 nuclear genes encoding mitochondrial proteins, diagnosing these disorders can be challenging. Mitochondrial disorders associated with mutations in nuclear DNA can involve the respiratory chain, coenzyme Q10 biosynthesis, mitochondrial protein synthesis, or mtDNA maintenance. The common clinical phenotypes associated with mitochondrial disorders are poor growth, myopathy, exercise intolerance, developmental delay, neurological problems, visual and/or hearing problems, diabetes, and heart, liver, or kidney disease.

The overall prevalence of mitochondrial disorders is expected to be at 1:5000 affected individuals. A mitochondrial disorder evaluation is often warranted in children with neurologic dysfunction with multisystem involvement. The presence of distinctive phenotypes coupled with family history and genetic testing can make the diagnosis of a mitochondrial disorder straightforward. However, when a patient does not have a distinctive phenotype, measurement of plasma and CSF lactate, acylcarnitines, and urine organic acids can be helpful. Direct measurement of oxidative phosphorylation by high-resolution respirometry is another emerging technique that may be useful in confirming the presence of mitochondrial disease. Genomic testing can include sequencing of the exome, genome, or mtDNA. However, the main limitations of genetic approaches include false negatives due to poor sequence coverage (single gene/multiple

panel testing can be more accurate in these cases) and inability to detect deletions, duplications, triple repeat expansions, and epigenetic modifications.

Glycogen storage diseases

Glycogen provides a source of glucose when the glucose is not supplied continuously. Metabolism of glycogen is tightly regulated by kinases and phosphatases that activate and inactivate glucose-6-phosphatase and its kinase during periods of fasting in response to increased glucagon. Glycogen storage diseases (GSDs) comprise a large category of disorders caused by abnormalities in both the degradation and synthesis of glycogen. They present with either hypoglycemia or symptoms of organ dysfunction due to the accumulation of glycogen. Some GSDs have predominantly hepatic-associated issues with hepatomegaly, whereas others have predominant muscle presentations with associated muscle cramps and exercise intolerance. Forearm exercise testing can be used to identify the myopathic forms of GSDs, because lactic acid production during exercise is minimal in the absence of glucose release from glycogen. Similar results are obtained in glycolytic defects, because released glucose cannot be metabolized to pyruvic acid. Treatment is aimed at the prevention of hypoglycemia in those GSDs associated with short fast hypoglycemia.

Deficiency of acid maltase, also known as Pompe disease, is unique among the GSDs in that glycogen accumulates in the lysosome. Severe acid maltase deficiency results in hypotonia and hypertrophic cardiomyopathy in early infancy. Milder deficiency results in muscle weakness, typically associated with elevation of creatine kinase. Diagnosis of Pompe disease is usually based on enzymatic analysis of either liver or dried blood spots. Excretion of glucose tetrasaccharides in urine is increased in infants with Pompe disease, and quantitation of this metabolite can be used to monitor the effectiveness of enzyme replacement therapy.

Peroxisomal disorders

Peroxisomes are important in the metabolism of branched-chain, polyunsaturated, and very long-chain fatty acids. They also play a role in the synthesis of bile acids, isoprenoids, and ether phospholipids, such as plasmalogens. Peroxisomal disorders can arise due to either defects in peroxisome biogenesis or single peroxisomal enzyme deficiencies. Peroxisome biogenesis disorders (PBDs) are a group of conditions that result in deficiency of multiple peroxisomal functions due to inability to make peroxisomes or deficiency in the import of intraperoxisomal proteins. PBDs include the Zellweger spectrum disorders (Zellweger syndrome, neonatal adrenoleukodystrophy, and infantile Refsum disease) and rhizomelic chondrodysplasia punctata type 1. The most severe forms of these conditions present in the neonatal period with severe hypotonia, seizures, liver and renal disease, calcific stippling of long bone epiphyses, and dysmorphic facial features resembling Down syndrome.

Deficiencies of single peroxisomal enzymes have also been described. The most common of these are X-linked adrenoleukodystrophy (XALD) and Refsum disease. XALD is due to the deficiency of a peroxisomal membrane fatty acid transporter, which results in progressive neurologic deterioration and dementia starting in the early school years in the most severe form. XALD is a common cause of adrenal insufficiency in boys. A milder variant, adrenomyeloneuropathy, presents with increased spasticity in adulthood and may be misdiagnosed as hereditary spastic paraparesis. Some female carriers of XALD develop gait abnormalities in later adulthood. Although the diagnosis of XALD is possible based on measurement of very long-chain fatty acid (VLCFA), this test cannot predict the phenotype and is only abnormal in 80% of carrier females.

In Refsum disease, deficiency of phytanoyl-CoA hydroxylase results in accumulation of phytanic acid with resulting retinitis pigmentosa, sensory and motor neuropathies, and ataxia. Onset of symptoms most commonly occurs in late childhood or adolescence. Night blindness and retinitis pigmentosa are often the first signs of Refsum disease. Because phytanic acid is obtained via exogenous sources, a phytanic acid-restricted diet is effective in preventing progression of symptoms if the diet is started early enough. Of note, adult-onset Refsum disease is distinct from infantile Refsum disease, which is a PBD within the Zellweger spectrum.

Measurement of VLCFA, pipecolic acid, and phytanic acid in plasma and plasmalogens in red blood cells are commonly used to screen for peroxisomal disorders. Currently, XALD is the only peroxisomal disorder included on the RUSP. Screening for XALD typically involves the measurement of VLCFA, C26, by tandem mass spectrometry. A plasma fatty acid profile, performed by GC−MS, which includes measurement of VLCFAs (C22, C24, and C26), phytanic acid and pristanic acid can aid in evaluating peroxisomal disorders. In many cases, molecular analysis of peroxisomal genes allows for definitive diagnosis and identification of at-risk family members.

Lysosomal storage diseases

Lysosomes are responsible for degradation of a variety of cellular waste products. A number of IEMs have been described that result in accumulation of various materials in the lysosome. Features that suggest a lysosomal storage disease (LSD) include hepatosplenomegaly, coarsening

facial features, bony changes indicative of dysostosis multiplex, developmental regression, corneal clouding, and nonimmune hydrops fetalis. In addition, some LSDs may now be identified via newborn screening using multiplexed LC−MS/MS-based methods. LSDs are divided into mucopolysaccharidoses (MPSs), oligosaccharidoses, and sphingolipidoses.

MPSs occur when the degradation of the carbohydrate moiety of proteoglycans that are present at cell surfaces and in the extracellular matrix is incomplete due to deficiency of a lysosomal enzyme. The glycosaminoglycans (GAGs) that accumulate can be quantitated in urine. Separation of different classes of GAGs by thin-layer chromatography, electrophoresis, or HPLC allows identification of the accumulating GAGs and helps determine the most appropriate enzyme assay for definitive diagnosis. In some MPSs, the total excretion of GAGs is only mildly increased and separation of the GAGs is necessary to recognize the accumulation of a particular GAG. This is commonly seen with Sanfillippo syndrome types A−D and Morquio syndrome types A and B. In both cases, the disease subtypes are clinically and biochemically indistinguishable through GAG analysis alone; however, diagnosis can be made by either enzymatic or molecular testing. Development of recombinant enzyme replacement therapy has expanded the treatment of MPSs, but the infused enzyme does not cross the blood−brain barrier and the treatment is not effective against the neurodevelopmental effects of these conditions.

Individuals with oligosaccharidoses have abnormalities in the degradation of the oligosaccharide groups of glycoproteins. Their clinical features are similar to those seen in MPSs, but organomegaly is less common and measurement of GAGs is normal. Vacuolated lymphocytes may be noted in many of the oligosaccharidoses. Thin-layer chromatography or HPLC of oligosaccharides in urine reveals the accumulation of abnormal material that must be identified by comparison to the pattern observed in individuals with confirmed oligosaccharidoses. Newer methods for oligosaccharide analysis, which employ the use of matrix-assisted laser desorption/ionization-TOF mass spectrometry to analyze permethylated urine oligosaccharides, have been described and can be used for screening individuals with suspected oligosaccharidosis. No treatment is currently available for oligosaccharidoses.

Sphingolipids are complex lipids containing the amino alcohol sphingosine attached to a long-chain fatty acid with or without other polar groups or sugars. Many are required for myelin, and gangliosides (a sialic acid containing glycosphingolipid) are important in gray matter. Enzyme deficiencies have been described for each of the enzymes required to degrade sphingolipids. Some sphingolipidoses are associated predominantly with hepatosplenomegaly; these include Gaucher disease and Niemann−Pick disease types A and B. A more severe form of Gaucher disease results in neurodegenerative disease and may be due to more severe enzyme deficiency, because both forms are due to deficiency of glucocerebrosidase.

Fabry disease, an X-linked condition, results in storage of glycolipids in reticuloendothelial cells, which results in episodes of neuropathic pain and eventually in renal insufficiency. One form of Fabry disease results in hypertrophic cardiomyopathy. Metachromatic leukodystrophy and globoid (Krabbe) leukodystrophy are sphingolipidoses associated with prominent demyelination and peripheral neuropathy; white matter changes are prominent on MRI. The gangliosidosies, including Tay−Sachs disease, are characterized by neuronal storage that can be seen as a cherry red spot in the infantile presentation. In those sphingolipidoses that result in neurologic abnormalities, early onset disease typically presents with developmental regression, whereas psychiatric disturbances are more common with adult onset. Diagnosis of these conditions requires measurement of enzyme activity in leukocytes or fibroblasts, but the results must be interpreted with caution. In rare cases of gangliosidosis, enzyme activity is normal in vitro, but deficiency of the GM2-activator protein results in enzyme deficiency in vivo. In other cases, apparent enzyme deficiency in an asymptomatic adult is actually a pseudodeficiency due to deficient activity against the artificial substrate. Finally, in I-cell disease, which results from impaired targeting of lysosomal enzymes to their proper intracellular location, multiple enzymes may show increased activity in a plasma (but not cellular) sample due an increased concentration.

Congenital disorders of glycosylation

Congenital disorders of glycosylation are a continuously expanding group of disorders that affect glycoprotein biosynthesis. Protein glycosylation is an important post-translational modification that impacts protein folding, subcellular localization, as well as antigen−antibody and protein−receptor interactions. A variety of sugar chains are covalently attached to proteins by either O-glycosidic linkage to serine or threonine residues or by *N*-linkage to amino groups of asparagine residues. Measurement of carbohydrate-deficient transferrin is often used to screen for congenital disorders of glycosylations (CDGs) with deficient *N*-glycosylation. This testing was initially done by isoelectric focusing, but the use of this test as a screen for chronic alcohol abuse has resulted in the development of turbidimetric immunoassay, capillary zone electrophoresis, HPLC, and LC-MS methods that are more readily available. For some of these testing methods, interpretation of results is complicated by some benign variants

of the transferrin protein that alter the electrophoretic pattern.

Type 1 CDGs affect synthesis of the carbohydrate before transfer to the protein, whereas conditions affecting maturation of the carbohydrate group are classified as type 2. Within each type, disorders are lettered in order of report. The most common CDG is type PMM2-CDG (1a), which is due to deficiency of phosphomannomutase; this deficiency prevents conversion of mannose 6-phosphate to mannose 1-phosphate, which is required for transfer of mannose to GDP. This deficiency subsequently prevents the transfer of mannose to the carbohydrate backbone. Although many of these mannose residues are removed during later maturation of the glycoprotein, they are required for glycoprotein processing, and their absence prevents further modification of the carbohydrate group. Individuals with PMM2-CDG typically have unusual fat distribution with inverted nipples, coagulation abnormalities, cerebellar hypoplasia, developmental delay, and liver function abnormalities; many die in infancy, although adults have been reported. Most CDGs are associated with significant developmental disabilities, but gastrointestinal symptoms (diarrhea, liver fibrosis, and failure to thrive) predominate in types 1b and 1h; intellectual development is normal in these variants. Phosphomannose isomerase deficiency (CDG1b) is the only treatable CDG, and symptoms are decreased by treatment with mannose. In addition to genetic testing, CGDs with impaired *N*-glycosylation can be distinguished by measuring the relative abundance of specific *N*-glycans using LC−MS/MS.

Conclusion

IEMs are individually rare, yet collectively common. Since severe morbidity can be avoided in some diseases, it is important to make the correct diagnosis as soon as possible. The role of the clinical laboratorian in the diagnosis of IEM cannot be underestimated. Diagnosis of IEMs relies on specialized laboratory testing and its interpretation. It is important to rule out diseases that can be treated first, and to perform diagnostic testing before confirmatory or prognostic genetic testing.

Acknowledgment

The authors gratefully acknowledge the original contributions of Cheryl Garganta, Stanley Lo, and Sarah Brown upon which portions of this chapter are based.

References

[1] D.J. Dietzen, A.L. Weindel, M.O. Carayannapoulos, E.T. Landt, T. E. Reimschisel, C.H. Smitch, Rapid comprehensive amino acid analysis by liquid chromatography/tandem mass spectrometry: comparison to cation exchange with post-column ninhydrin detection, Rapid Commun. Mass. Spectrom. 22 (22) (2008) 3481−3488.

[2] J. Mak, T.M. Cowan, A. Le, Quantitative analysis of underivatized amino acids by liquid chromatography-tandem mass spectrometry, Methods Mol. Biol. Amino Acid Anal. 22 (22) (2019) 85−109.

[3] J.D. Sharer, I. De Biase, D. Matern, S. Young, M.J. Bennett, A.A. Tolun, et al., Laboratory analysis of amino acids, 2018 revision: a technical standard of the American College of Medical Genetics and Genomics (ACMG), Genet. Med 20 (12) (2018) 1499−1507.

[4] J.D. Shoemaker, W.H. Elliot, Automated screening of urine samples for carbohydrates, organic and amino acids after treatment with urease, J. Chromatogr. 562 (1−2) (1991) 125−138.

[5] J. Vockley, H. Anderson, K. Antshel, N. Braverman, et al., Phenylalanine hydroxylase deficiency: diagnosis and management guideline, Genet. Med. 16 (2) (2014) 188−200.

[6] A. Kumps, O. Duez, Y. Mardens, Metabolic, nutritional, Iatrogenic, and artifactual sources of urinary organic acids: a comprehensive table, Clin. Chem. 48 (5) (2002) 708−717.

Further readings

N. Blau, M. Duran, M.E. Blaskovics, K.M. Gibson (Eds.), Laboratory Guide to the Methods in Biochemical Genetics, Springer-Verlag, New York, 2008.

U. Garg, L.D. Smith, B.A. Heese (Eds.), Laboratory Diagnosis of Metabolic Diseases,, AACC Press, Washington, DC, 2012.

J.M. Saudubray, G. Van den Berghe, J.H. Walter, Inborn metabolic diseases: diagnosis and treatment, Fifth ed., Springer Verlag, Berlin, 2012, p. 561.

Self-assessment questions

1. Most amino acid disorders are identified in which matrix?
   a. serum
   b. plasma
   c. urine
   d. CSF
2. Organic acid analysis most commonly uses which instrument?
   a. HPLC
   b. MS/MS
   c. LC−MSMS
   d. GC−MS
3. Acylcarnitine analysis is most commonly used to identify which class of disorders?
   a. urea cycle disorders
   b. glycogen storage disorders
   c. lysosomal storage disorders
   d. fatty acid oxidation disorders
4. Which amino acid is pathognomonic for MSUD?
   a. valine
   b. leucine
   c. isoleucine
   d. alloisoleucine
5. What is the most common fatty acid oxidation disorder?
   a. VLCAD
   b. LCAD
   c. MCAD
   d. SCAD

Answers

1. b
2. d
3. d
4. d
5. c

Chapter 49

The porphyrias: fundamentals and laboratory assessment

Elizabeth L. Frank[1,2] and M. Laura Parnás[3]

[1]Department of Pathology, University of Utah School of Medicine, Salt Lake City, UT, United States, [2]ARUP Laboratories, Inc., Salt Lake City, UT, United States, [3]Roche Diagnostics Corporation, Indianapolis, IN, United States

Introduction

Porphyrias are a group of rare metabolic disorders associated with defects in the heme biosynthetic pathway [1]. Partial deficiency or hyperactivity of one of the heme biosynthesis enzymes results in excess pathway intermediates that accumulate in body tissues and are excreted. These precursor compounds cause distinctive biochemical features and disease symptoms. Porphyria symptoms can mimic a variety of conditions, and thus porphyria can be difficult to diagnose.

Traditional classification of the porphyrias as hepatic or erythropoietic is based on the primary tissue in which overproduction of heme precursors occurs; however, porphyrias can be divided into two groups, acute (neurologic) and cutaneous (nonacute), according to their dominant clinical manifestations [2]. Characteristic patterns of accumulation of heme precursors due to abnormal enzyme activities are associated with each porphyria. These compounds can be detected readily in blood, urine, and feces, which permits biochemical assessment [3,4]. Appropriately chosen and interpreted, laboratory tests for porphyria are sensitive and specific, and play a crucial role in the diagnosis and management of these disorders.

Biochemistry: heme formation

Biosynthesis of heme occurs in all nucleated cells, primarily in developing erythrocytes of the bone marrow in which hemoglobin is generated, and to a lesser extent in hepatocytes, the site of formation of heme-containing enzymes [1]. The tetrapyrrolic heme structure is generated enzymatically in a sequential series of steps that takes place in two distinct cellular compartments; the first and the last three steps occur in the mitochondrion, and the intermediate steps take place in the cytosol (Fig. 49.1).

Eight enzymes catalyze the formation of protoporphyrin IX and the chelation of iron to produce heme. The initial and rate-limiting reaction in the pathway is the condensation of glycine and succinyl-coenzyme A to form δ-aminolevulinic acid (ALA), which is catalyzed by δ-aminolevulinic acid synthase (ALAS), Enzyme Commission (EC) number 2.3.1.37 [5]. Condensation of two molecules of ALA to generate the monopyrrole porphobilinogen (PBG) is mediated by PBG synthase (alternate name ALA dehydratase; EC 4.2.1.24). This step is followed by polymerization of four molecules of PBG to form the linear tetrapyrrole hydroxymethylbilane. The reaction is catalyzed by hydroxymethylbilane synthase (HMBS, alternate name PBG deaminase; EC 2.5.1.61); however, spontaneous polymerization can occur if PBG concentrations are high. Cyclization of the tetrapyrrole to form uroporphyrinogen III is catalyzed by uroporphyrinogen III synthase (EC 4.2.1.75) [6], although minor uncatalyzed cyclization occurs and the structural isomer uroporphyrinogen I is generated. Decarboxylation of uroporphyrinogen isomers is mediated stepwise through uroporphyrinogen decarboxylase (UROD; EC 4.1.1.37) to form the tetracarboxyl isomers coproporphyrinogen III and I. Only coproporphyrinogen III is oxidatively decarboxylated by coproporphyrinogen oxidase (CPOX; EC 1.3.3.3) to protoporphyrinogen IX [1]. Subsequent oxidation of the compound in a reaction catalyzed by protoporphyrinogen oxidase (PPOX; EC 1.3.3.4) produces protoporphyrin IX, and insertion of ferrous iron (Fe^{2+}) by ferrochelatase (FECH; EC 4.99.1.1) generates heme.

Heme biosynthesis is regulated in hepatic parenchymal cells by negative feedback inhibition. The intracellular concentration of free heme (heme pool) regulates synthesis of the ubiquitous ALAS isoform ALAS1 to control production of heme by inhibiting the first step of the

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00049-1

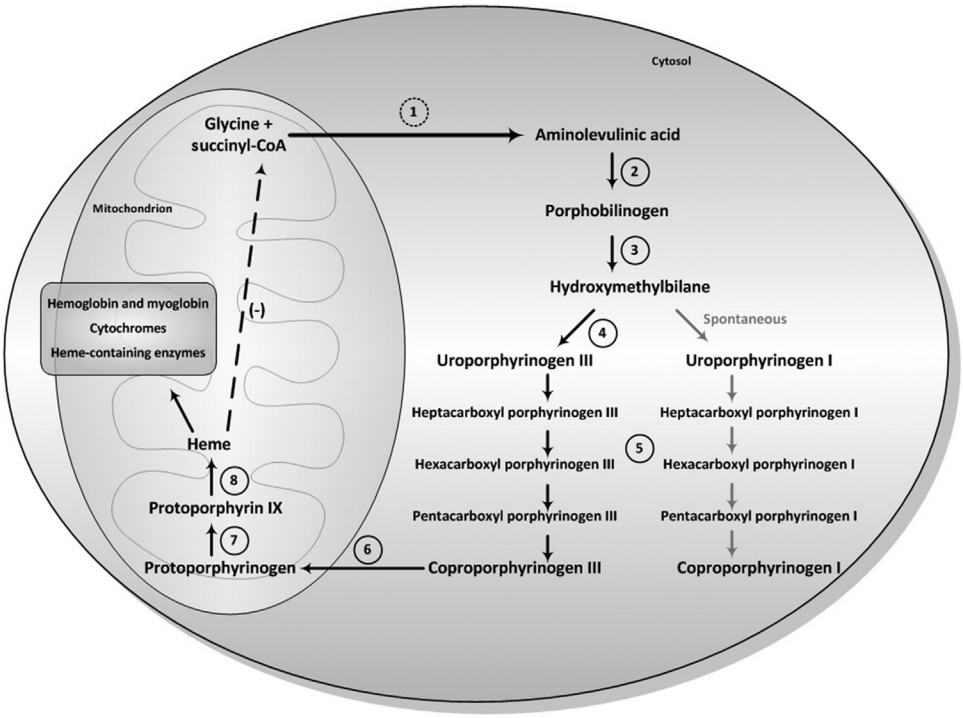


FIGURE 49.1 Heme biosynthetic pathway. Eight enzymes catalyze synthesis of heme from glycine and succinyl-coenzyme A. Heme production controls the hepatic pathway by regulating aminolevulinic acid synthase (1) activity. The other enzymes are: porphobilinogen synthase (2), hydroxymethylbilane synthase (3), uroporphyrinogen III synthase (4), uroporphyrinogen decarboxylase (5), coproporphyrinogen oxidase (6), protoporphyrinogen oxidase (7), and ferrochelatase (8). Production of heme regulates the activity of aminolevulinic acid synthase, the rate-limiting enzyme in the pathway, by negative feedback inhibition (−). Heme synthesis in erythroid cells is limited by iron (Fe^{2+}) availability. *Adapted with permission from M.L. Parnas, E.L. Frank, Porphyrias: a guide to laboratory assessment, Clin. Lab. News 36(4) (2010) 8−10 (Copyright AACC, used with permission).*

pathway. In erythropoietic cells, heme biosynthesis is controlled by the erythroid-specific enzyme ALAS2. Synthesis of ALAS2 is induced during active heme biosynthesis, and the rate of heme production is limited by the availability of iron (Fe^{2+}) [7].

Under usual conditions, the highly efficient biosynthetic pathway allows for most of the ALA produced to be converted to heme. Only minimal amounts of precursors and intermediates accumulate and are excreted, with the route of elimination directed largely by the intrinsic aqueous solubility of each compound. The water-soluble porphyrin precursors ALA and PBG are excreted in urine. The octacarboxylic intermediate uroporphyrinogen is water-soluble and excreted renally; the hydrophobic dicarboxylated protoporphyrin IX is excreted in feces via the biliary tract. The remaining porphyrin intermediates, including the tetracarboxylated coproporphyrinogens, are somewhat water-soluble and appear in both urine and feces. The porphyrinogens, cyclic tetrapyrrole intermediates that exist in the body in a reduced state, are oxidized rapidly on exposure to air to corresponding porphyrin compounds, the analytes measured in the clinical laboratory [2]. The precursors and porphyrins found in blood, plasma, urine, and feces are shown in Table 49.1.

Genetic disorders: porphyrias

Genetic abnormalities in any of the eight heme biosynthetic enzymes give rise to the porphyrias [1]. Inheritance for most of these diseases is autosomal dominant, and one functional gene is present. Enzyme deficiency is partial, and activity is sufficient to maintain heme homeostasis. In certain clinical circumstances, pathway metabolites that precede the enzymatic defect accumulate in the body, causing characteristic signs and symptoms of the various porphyrias. An X-linked dominant gain-of-function mutation in the erythroid-specific ALAS2 enzyme promotes enzyme activity and production of excess protoporphyrin to cause an X-linked dominant protoporphyria [8]. Although each type of porphyria originates from a different genetic abnormality, the clinical manifestations of these disorders are similar, allowing their classification into acute and cutaneous porphyrias [2].

Acute porphyrias

The disorders comprising this group of porphyrias are acute intermittent porphyria (AIP), variegate porphyria (VP), hereditary coproporphyria (HCP), and ALA dehydratase

TABLE 49.1 Accumulation and excretion patterns of heme precursors and porphyrin intermediates in body fluids during symptomatic episodes of porphyria. Diseases are grouped into acute and cutaneous porphyrias and listed in order of decreasing prevalence.

Porphyria	Enzyme	Erythrocyte porphyrins	Plasma fluorescence (nm)	Urine precursors	Urine porphyrins	Fecal porphyrins
Acute porphyrias						
Acute intermittent porphyria	HMBS	Not increased	618–622	PBG (ALA)	Uro (from PBG)	Not increased
Variegate porphyria	Protoporphyrinogen oxidase	Not increased	626–628	PBG (ALA)	Uro, copro III	Copro III, proto IX
Hereditary coproporphyria	Coproporphyrinogen oxidase	Not increased	618–622	PBG (ALA)	Copro III, uro	Copro III
ALA dehydratase deficiency porphyria	Porphobilinogen synthase	Zn-proto	N/A	ALA	Copro III	Not increased
Cutaneous porphyrias						
Porphyria cutanea tarda	Uroporphyrinogen decarboxylase	Not increased	618–622	Not increased	Uro, hepta, and isocopro	Isocopro
Erythropoietic protoporphyria	Protoporphyrin ferrochelatase	Proto IX	634–636	Not increased	Not increased	Proto IX
X-linked erythropoietic protoporphyria	ALA synthase (ALAS2)	Free proto IX; Zn-proto	634–636	Not increased	Not increased	Proto IX
Congenital erythropoietic porphyria	Uroporphyrinogen III synthase	Uro I Copro I	618–622	Not increased	Uro I, copro I	Copro I

ALA, Aminolevulinic acid; *ALAS*, aminolevulinic acid synthase; *Copro*, coproporphyrin; *Hepta*, heptacarboxyl porphyrin; *HMBS*, hydroxymethylbilane synthase; *Isocopro*, isocoproporphyrin; *PBG*, porphobilinogen; *Proto IX*, protoporphyrin IX; *Uro*, uroporphyrin; *Zn-proto*, zinc protoporphyrin.
Source: Used with permission from E.L. Frank, M.L. Parnas, The porphyrias: fundamentals and laboratory assessment, in: W. Clarke (Ed.), Contemporary Practice in Clinical Chemistry, third ed., AACC Press, Washington, DC, 2016, pp. 665–671.

deficiency porphyria (ADP). Each disorder is associated with a separate enzymatic defect (Table 49.1). AIP, VP, and HCP are inherited in an autosomal dominant manner; the extremely rare ADP is inherited as an autosomal recessive trait. AIP is the most common of the acute porphyrias, affecting approximately 1:75,000 Caucasian persons [4]. Acute porphyrias are characterized by acute, potentially life-threatening neurovisceral attacks with symptoms of diffuse abdominal pain, gastrointestinal ailments, peripheral neuropathies, and mental disturbances [2]. Hypertension and tachycardia are typical. Hyponatremia and seizures may occur during attacks [3]. Acute symptomatic episodes typically commence during early adulthood, occur more frequently in women than men, and may be accompanied by skin lesions in VP and HCP. Pediatric onset is unusual, but has been documented [9].

The genetic abnormalities present in the three autosomal dominant acute porphyrias lead to an enzymatic activity approximately 50% of that expected. Adequate heme concentration is maintained as a result of upregulation of ALA synthase. Consequently, the acute porphyrias have low clinical penetrance, with most individuals remaining asymptomatic for life and approximately 10% of persons with the genetic defect manifesting symptoms in the presence or absence of precipitating endogenous and exogenous factors [4]. These exacerbating agents include drugs, hormones, infection, or stress, and act by increasing hepatic demands for heme, which induces heme synthesis. The enzyme deficiency becomes rate-limiting, and precursors and intermediates accumulate. Incidence and severity of the attacks are variable and more frequent upon exposure to precipitating factors. An increase in urinary excretion of PBG and/or ALA is observed in all patients with acute symptoms and in some asymptomatic individuals [10]. In ADP, symptoms are precipitated by excess ALA, and only urinary ALA is elevated.

Insufficiency of PPOX (VP) or CPOX (HCP) produces porphyrins that inhibit HMBS, the third enzyme in the pathway, and consequently elevate PBG concentrations. For this reason, both VP and HCP are categorized as acute porphyrias; however, photosensitivity and skin lesions are the only clinical manifestation in approximately 60% of individuals with VP and some (~5%) of those with HCP [4].

Cutaneous porphyrias

The cutaneous porphyrias include porphyria cutanea tarda (PCT), erythropoietic protoporphyria (EPP), X-linked erythropoietic protoporphyria (XLP), and congenital erythropoietic porphyria (CEP) [11]. These disorders and the associated enzymes are listed in Table 49.1. The most prevalent nonacute porphyria, PCT, can be inherited as an autosomal dominant trait, but usually the disorder is sporadic in origin due to acquired inhibition of UROD activity in hepatocytes [12]. Approximately one person in 10−20,000 is affected. Excess alcohol consumption, use of tobacco or estrogens, hepatitis C or human immunodeficiency viral infection, halogenated aromatic hydrocarbon exposure, and hemochromatosis or iron overload contribute to inactivation of hepatic UROD and are risk factors for PCT [12]. The mode of inheritance of EPP is complex, and frequently, disease expression is associated with a combination of two molecular defects [4]. XLP is due to an X-linked dominant gain-of-function mutation that causes hyperactivity of the erythroid-specific ALAS2 enzyme and excessive production of protoporphyrin [13]. Approximately 90% of protoporphyria cases are due to loss-of-function mutations in the enzyme FECH; 5% of cases result from activating mutations in ALAS2. The remainder are of unknown genetic basis [14]. The very rare CEP is autosomal recessive.

Clinically, the cutaneous porphyrias are characterized by photosensitization of the skin resulting from accumulation of porphyrin intermediates in tissues. Manifestations of disease are attributable to the light-absorbing properties of the porphyrin ring [2]. Symptoms, particularly on the hands, forearms, neck, and face, arise with sun exposure of the skin. PCT and CEP are characterized by increased skin fragility and blistering skin lesions that heal slowly, leaving scars and hyperpigmentation or hypopigmentation. Hypertrichosis of the face, ears, and arms is common. Symptoms are more severe and appear earlier in CEP patients, who may experience disfiguring mutilation of light-exposed body parts, such as the nose, ears, and hands. Late onset CEP associated with myeloid malignancy has been reported in older adults [15]. In EPP and XLP, acute painful burning sensations, redness, and itching on sun-exposed areas occur as a result of accumulation of protoporphyrin in erythrocytes. Disease onset typically occurs in childhood.

In addition to these diseases, rare homozygous forms of porphyria, such as hepatoerythropoietic porphyria (HEP), and dual porphyrias attributed to the presence of two different heterozygous defects have been reported [1].

Diagnosis: laboratory evaluation

Although mutation analysis and enzyme assays are available, laboratory testing to diagnose porphyrin disorders usually commences with identification of characteristic patterns of excess porphyrin precursor and intermediate compounds in body fluids. The specimen selected for testing (blood, urine, or feces) depends on solubility of the metabolic intermediates. The porphyrin precursors ALA and PBG are water-soluble compounds that are excreted and detected in urine. Aqueous solubility of porphyrins varies with the number of carboxylic acid side chains attached to the tetrapyrrole ring. Uroporphyrin, with eight carboxylic acid groups, is soluble in aqueous solution and can be detected in plasma and urine. Hepta-, hexa-, and pentacarboxylated porphyrins, with seven, six, and five carboxyl groups, respectively, as well as the tetracarboxylated coproporphyrins, are measured in plasma, urine, and feces. Dicarboxylic porphyrins such as protoporphyrin are less water soluble, and are identified and measured in whole blood, plasma, and fecal specimens.

Porphyrin compounds are photoactive due to extensive conjugation of the tetrapyrrole ring, which allows absorption of light in the visible region [2]. Porphyrins absorb light of ~400-nm wavelength and emit a characteristic red fluorescence [16]. These photochemical properties produce some of the manifestations of porphyric disease and provide a means for detection of the compounds in body fluids tested in the clinical laboratory [17].

The goal of initial testing is to demonstrate abnormal amounts of porphyrin compounds. If excess porphyrins are observed, diagnosis is made by interpretation of the specific pattern of precursor and intermediate compounds in various body fluid specimens. Additional investigation can be performed to identify the enzymatic abnormality and guide treatment.

Initial testing for suspected porphyria

Testing for porphyria is guided by the clinical symptoms of the patient. For individuals who present with acute manifestations, including diffuse abdominal pain, peripheral neuropathy, and mental disturbances, and for whom there is high suspicion for an acute porphyria attack, measurement of PBG in a random urine specimen should be the first test ordered (Fig. 49.2).

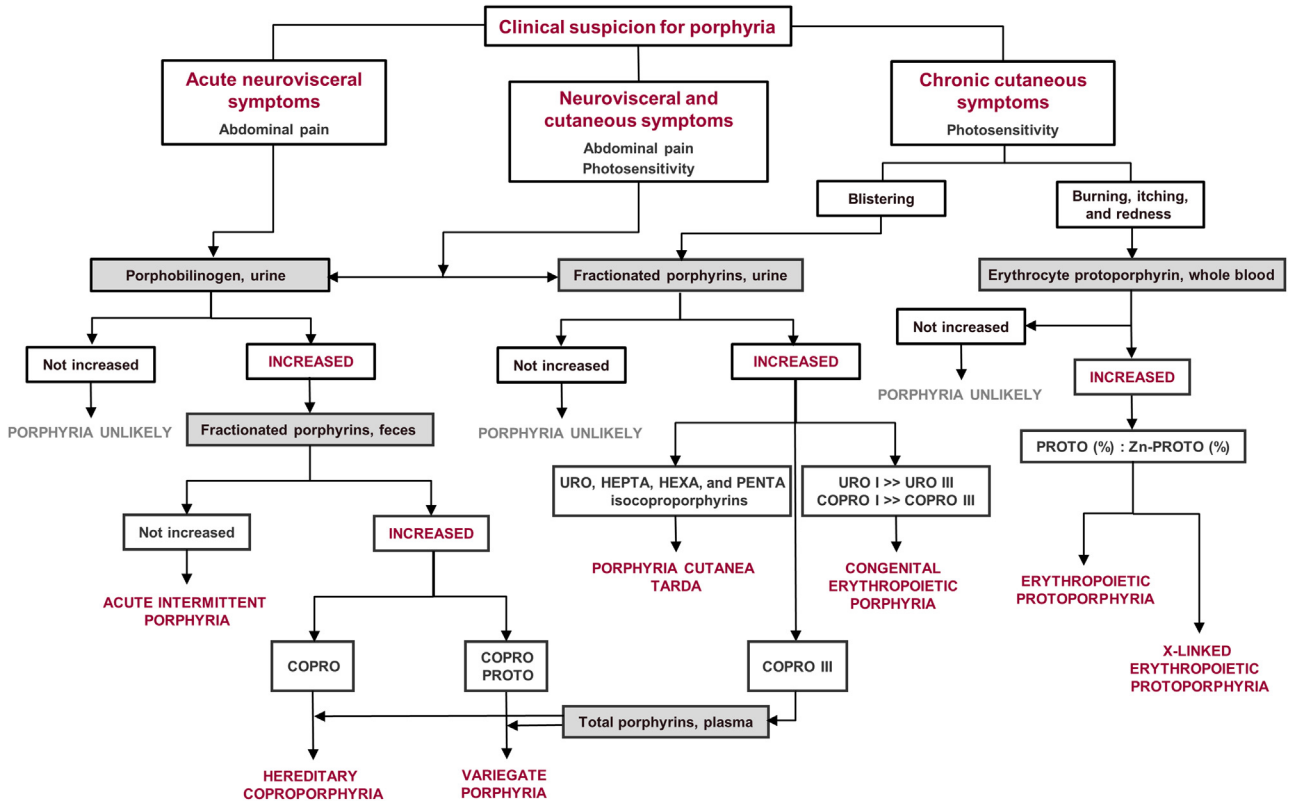


FIGURE 49.2 Diagnostic algorithm. Biochemical evaluation of the porphyrias. *COPRO*, Coproporphyrins; *COPRO I*, coproporphyrin I; *COPRO III*, coproporphyrin III; *HEPTA*, heptacarboxyl porphyrin; *HEXA*, hexacarboxyl porphyrin; *PENTA*; pentacarboxyl porphyrin; *PROTO*, protoporphyrin IX; *URO*, uroporphyrins; *URO I*, uroporphyrin I; *URO III*, uroporphyrin III; *Zn-PROTO*, zinc protoporphyrin. *Adapted with permission from M.L. Parnas, E.L. Frank, Porphyrias: a guide to laboratory assessment, Clin. Lab. News 36(4) (2010) 8–10 (Copyright AACC, used with permission).*

A common method for measurement of PBG utilizes ion-exchange chromatography to isolate the compound and remove interferences, followed by reaction with 4-dimethylaminobenzaldehyde (Ehrlich's reagent) to produce a rose red compound that is detected spectrophotometrically to determine PBG concentration [18]. Elevated PBG concentration is considered a positive result and strongly suggests the presence of an acute porphyria. Additional biochemical testing is used to elucidate the specific type of acute porphyria. Frequently ALA is elevated, although to a lesser extent than PBG, and, with the exception of identification of the extremely rare ADP, measurement of excess PBG is sufficient to diagnose acute porphyria. A negative PBG result in a urine specimen collected from a symptomatic patient effectively excludes an acute porphyria attack due to AIP, VP, or HCP. Measurement of ALA in urine is performed to exclude ADP. In the absence of symptoms, PBG may not be increased, and a negative result does not completely rule out acute porphyria.

Prompt diagnosis of a suspected attack of acute porphyria is desirable, but testing for PBG is not routine and, in many areas, is performed only in specialized laboratories. A semiquantitative rapid test for urinary PBG offered commercially is no longer available [19], and urgent testing may not be performed at a local laboratory. Other simple, fast laboratory-developed methods have been proposed as screening tests [20]. Assays for PBG using mass spectrometry and nuclear magnetic resonance (^{1}H-NMR) have been designed [21–23]. An ideal method would provide rapid and sensitive near-patient testing to support early initiation of therapy and avoid administration of potentially harmful (hemin) or inappropriate (opioid analgesics) treatment.

Recommendations in some guidelines specify evaluation of cutaneous symptoms using fluorescence scanning to screen plasma and urine specimens for total porphyrins [24]. This screening may be accomplished using chromatographic assays that produce quantitative results. Urinary porphyrin isomers can be separated and measured within 10–12 min by high-performance liquid chromatography (HPLC). Urine specimens require minimal sample preparation, and HPLC assays are more automated than spectroscopic methods used for total porphyrin testing. Employing a single assay rather than a qualitative screen followed by additional testing efficiently identifies porphyria and is particularly useful if porphyria testing is performed at distant specialized laboratories, since sequential screening and confirmation could delay diagnosis and treatment.

For adults with clinical manifestations of cutaneous photosensitivity (blisters and skin fragility), an appropriate test is quantitative fractionation of urinary porphyrins in a 24-h specimen (Fig. 49.2). Typically, individual porphyrins are separated by HPLC and detected using fluorescence [25]. The compounds fluoresce between 600 and 660 nm when dissolved in acidic solution [1]. Excess urinary excretion of porphyrins, particularly uroporphyrin and heptacarboxyl porphyrin, with the appearance of isocoproporphyrin, a metabolite produced in an alternate biosynthetic pathway in conditions of deficient UROD activity, is diagnostic of PCT [1]. Chromatographic separation of the nonfunctional porphyrin I isomers from porphyrin III isomers identifies CEP. Massive amounts of uroporphyrin I and coproporphyrin I isomers are found in the urine of individuals affected with this rare disorder.

HPLC analysis of urine specimens from individuals with HCP and VP can be expected to show increased concentrations of coproporphyrin III if the specimen was collected when cutaneous symptoms were present. This finding is not definitive; however, the results can be used to exclude PCT and may prompt testing for PBG if the clinical situation is consistent with acute porphyria. Urinary porphyrins may not be increased if cutaneous symptoms are absent. Although AIP is not associated with cutaneous symptoms, a specimen collected from an affected individual during an acute attack may reveal significantly increased uroporphyrin due to spontaneous polymerization of PBG in the specimen. If PBG analysis was not ordered, the specimen should be tested for PBG to detect AIP. If urinary porphyrins are not increased, porphyria probably is not the cause of observed symptoms.

Individuals who experience acute phototoxic reaction to sun exposure (pain, redness, and itching of the face and hands or feet), particularly children, should be screened for elevated erythrocyte protoporphyrin in a whole blood specimen. Total measured porphyrins include both metal-free protoporphyrin and zinc protoporphyrin (ZPP), a metabolite formed by complexation of zinc ion with protoporphyrin [12]. Determination of ZPP as a fraction of total erythrocyte protoporphyrin distinguishes XLP from EPP [11,12]. Increased free erythrocyte protoporphyrin is consistent with EPP; in XLP, both free protoporphyrin and ZPP are elevated [4].

Additional biochemical testing

Following a positive screen for PBG and/or porphyrins, evaluation of plasma and fecal specimens provides additional diagnostic information. Quantitation of porphyrins in feces may be used to differentiate the acute porphyrias. Fecal porphyrin excretion is not increased appreciably in AIP. Significant fecal coproporphyrin excretion (10 or more times the upper reference limit) and an increased coproporphyrin III/I isomer ratio are characteristics of HCP, and VP is associated with increased fecal excretion of protoporphyrin and coproporphyrin [2]. Fecal isocoproporphyrin is a marker for PCT [26]. Testing plasma for porphyrin fluorescence confirms VP; a peak between 626 and 628 nm is pathognomonic [27]. Plasma fluorescence scanning can be used to monitor PCT and to support a diagnosis of protoporphyria.

HEP, a rare homozygous disorder of UROD deficiency, resembles CEP clinically, and can be differentiated by a predominance of protoporphyrin in blood and detection of isocoproporphyrin in feces [1].

Enzyme and molecular testing

Tests for enzymatic activity and molecular analysis, available from specialized laboratories [28], are best used to confirm a particular disorder in a previously diagnosed individual and to identify family members at risk for the disease [29]. Prediction of disease course or severity is not possible based on the results of molecular testing, but individuals carrying a mutation can be advised to prevent the development of symptoms by avoiding precipitating agents.

Secondary causes of increased porphyrin excretion

Abnormalities in metabolism and excretion of heme precursors and intermediates may occur in the absence of inherited porphyrias and can be caused by a variety of conditions (Table 49.2).

Isolated increases in concentrations of urinary coproporphyrin are encountered during porphyria testing and are typically associated with hepatic compromise similar to that associated with liver disease, drugs, or chronic consumption of alcohol [2]. In these conditions, coproporphyrin III is the predominant urinary isomer present. Impaired biliary excretion, such as that which occurs in obstructive jaundice, hepatitis, and cirrhosis, increases urinary coproporphyrin isomer I. Some inherited disorders of bilirubin metabolism (Dubin−Johnson and Rotor syndromes) alter biliary excretion of coproporphyrin isomers [30]. Urinary coproporphyrin may be increased, and more coproporphyrin I than coproporphyrin III is excreted renally. In unaffected individuals, the urinary coproporphyrin I/III isomer ratio is approximately 30%/70%. Inversion of the coproporphyrin isomer ratio to >80% coproporphyrin I is consistent with a diagnosis of Dubin−Johnson syndrome [31,32]. In Rotor syndrome, total urinary coproporphyrin and coproporphyrin I isomer concentrations are elevated; coproporphyrin I comprises approximately two-thirds of total coproporphyrin.

TABLE 49.2 Secondary causes of porphyrin metabolism abnormalities.

Condition	Abnormality
Hepatic dysfunction	
• Excess alcohol consumption and drugs	Urinary coproporphyrin III
Impaired biliary excretion	
• Cholestatic jaundice, hepatitis, and cirrhosis	Urinary coproporphyrin I
Disorders of bilirubin metabolism	
• Inherited jaundice	Urinary coproporphyrins I and III
Gastrointestinal system	
• Diet • Hemorrhage	Fecal porphyrins
Impaired renal function	
• Renal failure	Plasma porphyrins
Hematological disorders	
• Anemia	Zinc protoporphyrin
Heavy metal toxicity	
• Lead • Arsenic and mercury	ALA, coproporphyrin III, zinc protoporphyrin Urinary coproporphyrins
Hereditary tyrosinemia Type I	
• Enzyme inhibition by succinylacetone	ALA

ALA, Aminolevulinic acid.
Source: Modified from E.L. Frank, M.L. Parnas, The porphyrias: fundamentals and laboratory assessment, in: W. Clarke (Ed.), Contemporary Practice in Clinical Chemistry, third ed., AACC Press, Washington, DC, 2016, pp. 665–71 (Used with permission).

Lead intoxication increases porphyrins in body fluids and can cause attacks of acute abdominal pain and neurological disturbances that mimic acute porphyria attacks. Lead inhibits PBG synthase, the second enzyme in the heme biosynthesis pathway, and to a lesser extent CPOX [33]. As a result, lead toxicity is associated with significantly increased urinary ALA and coproporphyrin III excretion (Fig. 49.1). Lead reduces intracellular iron availability, and zinc replaces iron as a substrate for FECH; thus ZPP is formed in erythrocytes [34]. Tests for ALA, urinary porphyrins, and ZPP provide indirect evidence for lead toxicity. Definitive diagnosis is accomplished by demonstration of lead in blood and/or urine.

Specimen requirements

All specimens submitted for testing must be protected from light. Specimens should be stored and transported dark and cold. Porphyrins and precursor compounds are stable in unpreserved urine at 4°C for up to 48 h and can be frozen (−20°C) for several weeks [2]. Porphyrins in fecal specimens can be stored frozen (−20°C) for several months. Whole blood collected for protoporphyrin measurement should be anticoagulated with EDTA and stored in the dark at 4°C [35]. Urine specimens must be adequately concentrated to produce reliable results. Dilute urine specimens with creatinine concentrations of <25 mg/dL (~2 mmol/L) are not adequate for analysis. Creatinine concentration should be measured, and results for random urine collections should be reported as a ratio to creatinine concentration. Calculation using low creatinine concentrations measured in dilute urine will result in misleading analyte ratios.

Treatment and management of the porphyrias

Acute attacks of porphyria, which can be life-threatening, are treated by immediate withdrawal of suspected precipitating agents, treatment of coexisting illnesses and/or infections, and pain management with nonporphyrinogenic agents [4,36]. Treatment with intravenous heme preparations (hemin) is the preferred therapeutic approach. Hemin specifically inhibits hepatic ALAS activity, and effectively decreases urinary PBG and ALA

[3,4]. Medications that trigger porphyria attacks must be replaced with alternative medications [37]. Adequate nutrition and fluids must be sustained. Prevention of an acute attack is achieved through avoidance of precipitating factors, including specific drugs, alcohol, and tobacco; maintenance of adequate caloric intake, especially in the form of carbohydrates; and stress reduction [38]. Therapy can be monitored by measurement of urinary PBG excretion. Prophylactic hemin infusion may be considered, and liver transplant may be necessary for patients with severe, recurrent acute attacks. Potential new treatments include replacement enzyme therapy or gene therapy [39].

Treatment of nonacute porphyrias is primarily preventive. Cutaneous symptoms can be minimized by avoidance of ultraviolet (UV) and visible light exposure and use of protective clothing and topical opaque sunscreens [12,40]. For PCT, administration of low-dose chloroquine or hydroxychloroquine is used to stimulate hepatic secretion and urinary excretion of porphyrins. Therapeutic phlebotomy, recommended for patients with hemochromatosis, removes iron and decreases plasma ferritin concentration. Urinary or plasma porphyrin concentrations are measured to monitor disease. Precipitants such as excess alcohol consumption, use of hormones, iron supplements, and smoking should be avoided. Patients with EPP should avoid sun exposure through the use of protective clothing and topical sunscreens. Oral administration of β-carotene provides systemic photoprotection by quenching excited species formed by UV-activated porphyrins, thus preventing damage from oxidative radicals. A synthetic α-melanocyte-stimulating hormone analog, afamelanotide, may be used to promote melanin formation and enhance protective dermal pigmentation [41]. In CEP, strict avoidance of sunlight and protection of skin from trauma are essential. Blood transfusions help to decrease hemolysis and suppress overproduction of porphyrins [40]. Bone marrow transplantation is the only curative therapy for EPP and CEP.

Summary

The porphyrias are a group of rare diseases that result from genetic or acquired abnormalities in the enzymes required for heme biosynthesis. Porphyrias are characterized by accumulation of porphyrin precursor and intermediate metabolites that cause signs and symptoms of disease. Diagnosis of porphyrias is aided by laboratory testing to detect and quantify specific porphyrin precursors and intermediates in body fluids. Abnormalities in metabolism and excretion of heme precursors and intermediates may occur in the absence of porphyria. Conditions that alter porphyrin excretion include hepatobiliary malfunction, inherited disorders of bilirubin metabolism, and heavy metal intoxication.

References

[1] K.E. Anderson, S. Sassa, D.F. Bishop, R.J. Desnick, Disorders of heme biosynthesis: X-linked sideroblastic anemia and the porphyrias, in: D. Valle, A.L. Beaudet, B. Vogelstein, K.W. Kinzler, S.E. Antonarakis, A. Ballabio, K.M. Gibson, G. Mitchell (Eds.), The Online Metabolic and Molecular Bases of Inherited Disease, 2018. <http://ommbid.mhmedical.com> [accessed 30.11.18].

[2] M.N. Badminton, S.D. Whatley, A.C. Deacon, G.H. Elder, The porphyrias and other disorders of porphyrin metabolism, in: C.A. Burtis, E.R. Ashwood, D.E. Bruns (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, fifth ed., Elsevier Saunders, St. Louis, 2012, pp. 1031–1055.

[3] H. Puy, L. Gouya, J.-C. Deyback, Porphyrias, Lancet 375 (9718) (2010) 924–937.

[4] Z. Karim, S. Lyoumi, G. Nicolas, J.-C. Deybach, L. Gouya, H. Puy, Porphyrias: A 2015 update, Clin. Res. Hepatol. Gastroenterol. 39 (4) (2015) 412–425.

[5] ENZYME Nomenclature Database, <https://enzyme.expasy.org>; 2018 [accessed 04.12.18].

[6] A.R. Battersby, C.J. Fookes, G.W. Matcahm, E. McDonald, Biosynthesis of the pigments of life: formation of the macrocycle, Nature 285 (1980) 17–21.

[7] R.J. Desnick, M. Balwani, The porphyrias, in: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo (Eds.), Harrison's Principles of Internal Medicine, 20th ed., McGraw-Hill, New York, NY, 2019. <https://accessmedicine.mhmedical.com/> [accessed 03.01.19].

[8] S.D. Whatley, S. Ducamp, L. Gouya, B. Grandchamp, C. Beaumont, M.N. Badminton, et al., C-terminal deletions in the *ALAS2* gene lead to gain of function and cause X-linked dominant protoporphyria without anemia or iron overload, Am. J. Hum. Genet. 83 (3) (2008) 408–414.

[9] M. Balwani, P. Singh, A. Seth, E.M. Debnath, H. Naik, D. Doheny, et al., Acute intermittent porphyria in children: a case report and review of the literature, Mol. Genet. Metab. 119 (4) (2016) 295–299.

[10] A.K. Aarsand, P.H. Petersen, S. Sandberg, Estimation and application of biological variation of urinary delta-aminolevulinic acid and porphobilinogen in healthy individuals and in patients with acute intermittent porphyria, Clin. Chem. 52 (4) (2006) 650–656.

[11] M. Balwani, R.J. Desnick, The porphyrias: advances in diagnosis and treatment, Blood 120 (2012) 4496–4504.

[12] D.M. Bissell, K.E. Anderson, H.L. Bonkovsky, Porphyria, N. Engl. J. Med. 377 (2017) 862–872.

[13] M. Balwani, J. Bloomer, R. Desnick, X-linked protoporphyria, in: M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, K. Stephens, A. Amemiya (Eds.), GeneReviews, University of Washington, Seattle, Seattle(WA), 1993–2018.

[14] J.C. Whitman, B.H. Paw, J. Chung, The role of ClpX in erythropoietic protoporphyria, Hematol. Transfus. Cell Ther. 40 (2018) 182–188.

[15] P. Aguilera, C. Badenas, S.D. Whatley, J. To-Figueras, Late-onset cutaneous porphyria in a patient heterozygous for a uroporphyrinogen III synthase gene mutation, Br. J. Dermatol. 175 (2016) 1346–1350.

[16] C. Rimington, Spectral-absorption coefficients of some porphyrins in the Soret-band region, Biochem. J. 75 (1960) 620–623.

[17] M.B. Poh-Fitzpatrick, Molecular and cellular mechanisms of porphyrin photosensitization, Photodermatology 3 (1986) 148–157.

[18] D.J. Moore, R.F. Labbe, A quantitative assay for urinary porphobilinogen, Clin. Chem. 10 (12) (1964) 1105–1111.
[19] M. Vogeser, T. Stauch, Evaluation of a commercially available rapid urinary porphobilinogen test, Clin. Chem. Lab. Med. 49 (2011) 1491–1494.
[20] M. Roshal, J. Turgeon, P.M. Rainey, Rapid quantitative method using spin columns to measure porphobilinogen in urine, Clin. Chem. 54 (2) (2008) 429–431.
[21] R.E. Ford, M.J. Magera, K.M. Kloke, P.A. Chezick, A. Fauq, J.P. McConnell, Quantitative measurement of porphobilinogen in urine by stable-isotope dilution liquid chromatography-tandem mass spectrometry, Clin. Chem. 47 (9) (2001) 1627–1632.
[22] C.M. Benton, L. Couchman, J.T. Marsden, D.C. Rees, C. Moniz, C.K. Lim, Direct and simultaneous determination of 5-aminolaevulinic acid and porphobilinogen in urine by hydrophilic interaction liquid chromatography-electrospray ionization/tandem mass spectrometry, Biomed. Chromatogr. 26 (8) (2012) 1033–1040.
[23] M. Carichon, N. Pallet, C. Schmitt, T. Lefebvre, L. Gouya, N. Talbi, et al., Urinary metabolic fingerprint of acute intermittent porphyria analyzed by (1)H nmr spectroscopy, Anal. Chem. 86 (4) (2014) 2166–2174.
[24] J. Woolf, J.T. Marsden, T. Degg, S. Whatley, P. Reed, N. Brazil, et al., Best practice guidelines on first-line laboratory testing for porphyria, Ann. Clin. Biochem. 54 (2) (2017) 188–198.
[25] C.K. Lim, T.J. Peters, Urine and faecal porphyrin profiles by reversed-phase high-performance liquid chromatography in the porphyrias, Clin. Chim. Acta 139 (1984) 55–63.
[26] G.H. Elder, Differentiation of porphyria cutanea tarda symptomatica from other types of porphyria by measurement of isocoproporphyrin in faeces, J. Clin. Pathol. 28 (1975) 601–607.
[27] R.J. Hift, B.P. Davidson, C. Van der Hooft, D.M. Meissner, P.N. Meissner, Plasma fluorescence scanning and fecal porphyrin analysis for the diagnosis of variegate porphyria: Precise determination of sensitivity and specificity with detection of protoporphyrinogen oxidase mutations as a reference standard, Clin. Chem. 50 (5) (2004) 915–923.
[28] National Institutes of Health. Genetics Home Reference, <https://ghr.nlm.nih.gov/condition/porphyria#diagnosis>;2018 [accessed 30.11.18].
[29] S.D. Whatley, M.N. Badminton, Role of genetic testing in the management of patients with inherited porphyria and their families, Ann. Clin. Biochem. 50 (2013) 204–216. 29.
[30] M. Frank, M.O. Doss, Relevance of urinary coproporphyrin isomers in hereditary hyperbilirubinemias, Clin. Biochem. 22 (3) (1989) 221–222.
[31] J.R. Chowdhury, A.W. Wolkoff, N.R. Chowdhury, I.M. Arias, Hereditary jaundice and disorders of bilirubin metabolism, in: D. Valle, A.L. Beaudet, B. Vogelstein, K.W. Kinzler, S.E. Antonarakis, A. Ballabio, K.M. Gibson, G. Mitchell (Eds.), The Online Metabolic and Molecular Bases of Inherited Disease, 2018. <http://ommbid.mhmedical.com>; [accessed 30.11.18].
[32] N. Memon, B.I. Weinberger, T. Hegyi, L.M. Aleksunes, Inherited disorders of bilirubin clearance, Pediatr. Res. 79 (3) (2016) 378–386.
[33] R. Mamet, M. Sztern, A. Rachmel, B. Stahl, D. Flusser, N. Schoenfeld, Lead poisoning: a new biochemical perspective on the differentiation between acquired and hereditary neuroporphyria, Clin. Chem. 47 (2001) 1710–1713.
[34] R.F. Labbe, H.J. Vreman, D.K. Stevenson, Zinc protoporphyrin: a metabolite with a mission, Clin. Chem. 45 (1999) 2060–2072.
[35] A.C. Deacon, G.H. Elder, Best practice No. 165: front line tests for the investigation of suspected porphyria, J. Clin. Pathol. 54 (2001) 500–507.
[36] P. Stein, M. Badminton, J. Barth, D. Rees, M.F. Stewart, Best practice guidelines on clinical management of acute attacks of porphyria and their complications, Ann. Clin. Biochem. 59 (2013) 217–223.
[37] E. Pischik, R. Kauppinen, An update of clinical management of acute intermittent porphyria, Appl. Clin. Genet. 8 (2015) 201–214.
[38] 2018 Safe List. UK Porphyria Medicines Information Service (UKPMIS) and Cardiff Porphyria Service, <https://www.wmic.wales.nhs.uk/wp-content/uploads/2018/05/2018-Porphyria-safe-list-FINAL.pdf>; 2019 [accessed 10.01.2019].
[39] A. Fontanellas, M.A. Avila, P. Berraondo, Emerging therapies for acute intermittent porphyria, Expert. Rev. Mol. Med. 18 (2016) 1–13.
[40] S. Tintle, A. Alikhan, M.E. Horner, J.L. Hand, D.M.R. Davis, Cutaneous porphyrias part II: treatment strategies, Int. J. Dermatol. 53 (2014) 3–24.
[41] A.M. Lane, J.T. McKay, H.L. Bonkovsky, Advances in the management of erythropoietic protoporphyria – role of afamelanotide, Appl. Clin. Genet. 9 (2016) 179–189.

Self-assessment questions

1. Initial laboratory testing to evaluate symptoms of porphyria is based on _____.
   - **a.** analysis of DNA for mutations using genetic techniques
   - **b.** biochemical testing for porphyrins and precursor compounds
   - **c.** functional and quantitative enzyme assays
   - **d.** liquid chromatography with fluorescent detection
2. Acute porphyria can be identified by laboratory testing for _____.
   - **a.** abnormal heme biosynthesis
   - **b.** decreased activity of porphyrin decarboxylation enzymes
   - **c.** increased porphobilinogen in a random urine specimen
   - **d.** spontaneous production of functional porphyrin isomers
3. Cutaneous porphyrias can be diagnosed by _____.
   - **a.** characteristic patterns of porphyrin excretion in a urine specimen
   - **b.** identification of zinc protoporphyrin in a whole blood specimen
   - **c.** scanning fluorescence of a urine specimen
   - **d.** spectrophotometric testing of a random urine specimen
4. The resident on-call in the ER contacts the pathology lab to ask about testing for a 10-year-old male child with a rash on his cheeks that appears following sun exposure. Which test should the resident order?
   - **a.** quantitative porphobilinogen in a random urine specimen
   - **b.** fractionated urine porphyrins measured in a 24-h collection
   - **c.** fractionated porphyrins in a fecal specimen
   - **d.** free erythrocyte porphyrin in whole blood
5. Evaluation of a 4-year-old child with a stomachache who is irritable and cranky should include _____.
   - **a.** analysis of porphyrins in a random urine specimen
   - **b.** fecal porphyrin analysis
   - **c.** spectrophotometric analysis of urine for porphobilinogen
   - **d.** testing for lead in whole blood

Answers

1. b
2. c
3. a
4. d
5. d

Chapter 50

Basic pharmacokinetics

William Clarke[1] **and Mark A. Marzinke**[2]

[1]Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD, United States, [2]Departments of Pathology and Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- List the important processes involved in pharmacokinetics.
- Discuss factors that significantly influence pharmacokinetics in various populations.
- Perform basic pharmacokinetic calculations using standard equations.
- Discuss pharmacokinetic variance in special populations such as neonates or obese patients.

Introduction

The relationship between drug administration and drug response is a complicated process, and is dependent on a number of factors. These include (1) route of drug administration; (2) dosing frequency; (3) biological factors including renal and hepatic function; (4) drug−drug, drug−food, and drug−hormone interactions; (5) therapeutic regimen compliance; and (6) genetic variation. The relative contributions of each of these variables lead to a drug response postadministration. These concepts are encompassed in the field of pharmacology, which focuses on drug action, uses, and effects. At a fundamental level, pharmacology may be stratified into two parts—pharmacokinetics (PK) and pharmacodynamics (PD). PK is often thought of as "what the body does to the drug," while PD refers to the action of the drug, or "what the drug does to the body." Further, both of these processes may be affected by genetic disposition; pharmacogenetics, or the influence of genetic variation on PK and PD, is discussed in further detail in Chapter 53, Pharmacogenomics. The principles of PK are useful for the interpretation of laboratory results, but are particularly necessary in the areas of therapeutic drug monitoring (TDM) (see Chapter 51: Therapeutic drug monitoring) and clinical toxicology (Chapter 52: Toxicology and the clinical laboratory). While the PD of particular drugs can have significant impacts on laboratory testing, this field is broad and beyond the scope of this chapter.

Pharmacokinetics and influencing factors

Several biological processes influence drug pharmacology. Typically, PK encompasses the contributions of absorption, distribution, metabolism, and excretion (ADME) on drug kinetics. ADME represent the process of a drug getting into the body, traveling to a site of action where either biologic activity will be exerted or the drug is inactivated, and then, ultimately the drug or metabolite is removed from circulation. Variability in each of these processes can contribute to overall drug efficacy.

Drug dosing and administration

While the amount of drug administered (the dose) is certainly an important factor in PK, the amount that reaches the systemic circulation will vary significantly based on the route of administration. Drugs can be administered either intravascularly or extravascularly, and the route of drug delivery will be a significant driver of downstream PK and PD processes. Intravascular drugs are those that are administered directly into the veins (intravenous) or arteries (intraarterial) of the patient. There is an absorption phase when drugs are delivered intravascularly; drug effects (therapeutic or adverse) are often more immediate, and drug effects may be more potent. Further, there are few options for recourse postadministration, making potential adverse reactions difficult to control or reverse. Extravascular drugs can be introduced through a number of pathways including oral administration (tablet, capsule, and liquid), intramuscular (IM) injection (solution or suspension), subcutaneous (SQ) injection (solution or suspension), transdermal administration (patch), sublingual or buccal administration (tablet or film), or inhalation (metered inhaler). During extravascular drug delivery, additional considerations may include (1) drug absorption

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00050-8

from the site of administration into the general systemic circulation; (2) the time to onset of pharmacologic effect is dependent on drug formulation and subsequent release from its delivery modality; and (3) potential reduction in drug delivery to systemic circulation due to first-pass metabolism.

Liberation of the drug from its dosage formulation is a specific component of dose administration. Oral formulations are typically enteral, and are frequently prepared as tablets or capsules. Such delivery systems may contain fillers or binders and may be manufactured to modulate drug release kinetics; dosages may be timed, fast-acting, or sustained release, based on specific formulation components. Oral administration can also occur using parenteral (liquid) formulations. Parenteral formulations primarily involve drug preparation in a liquid or lyophilized matrix and is the most common formulation used in injectable drug delivery (such as intravenous, SQ, and IM administration). Drug stability and solubility are important considerations for these formulations. Finally, drugs can be formulated for topical application, and manufacturing processes typically involve emulsification between hydrophobic (i.e., oil) and water-based solutions. Due to the heterogeneity in both drug delivery systems and formulations, drug liberation can significantly affect the pharmacokinetic analysis of a drug and must be taken into account in various clinical scenarios.

In laboratory medicine, the primary consideration of pharmacokinetic analysis is to either assess potential toxic exposure or optimize therapy. As such, compliance (or adherence) with a prescribed regimen is a critical consideration when assessing PK. If a patient is not taking the drug as prescribed (or directed) or is taking substances that can alter the pharmacokinetic profile of the drug in question, this can affect the interpretation of test results and the pharmacologic assessment. Frequently, patients do not follow the directions for over-the-counter medications nor report herbal or nutritional supplements that can affect drug disposition [1]. Research has demonstrated that even with potential life-threatening conditions such as HIV infection or cancer, patients sometimes do not take their medication as directed [2,3]. In fact, one indication for TDM that is becoming more common is to assess compliance or adherence with prescribed drug regimens (see Chapter 51: Therapeutic drug monitoring).

Absorption

Absorption is the process of drug transport from the site of administration to the body compartment where it is usually measured, which is typically blood. As mentioned previously, intravenous drug administration circumvents this process, as drug is directly delivered into systemic circulation. For oral drug formulations, including tablets, capsules, and oral solutions, drugs may be absorbed by the gastric mucosa after dissolution or mixing with intestinal fluid. Within the gastrointestinal (GI) tract, the primary site for formulation dissolution and drug absorption is the small intestine. There are many factors that can affect drug absorption, including pH, drug formulation, and drug transporter activity. For example, there can be drug loss in the GI tract due to P-glycoprotein transport, gastric acid hydrolysis, or intestinal enzyme metabolism.

Following oral administration of a drug, first-pass metabolism frequently occurs, further contributing to the decreasing gradient in concentrations from drug administration to drug delivery to the systemic compartment. Once a drug is absorbed via the GI tract, it enters into the portal circulatory system prior to delivery to the systemic blood system. Absorbed drug is transported, via the portal vein, to the liver, where initial biotransformation can occur prior to export into the systemic circulation. Biotransformation may result in the activation or deactivation of a drug, and can therefore influence its local pharmacologic efficacy. Bioavailability is a primary PD driver for orally administered drugs and is discussed further below.

Notably, absorption is also a critical process for alternative routes of drug delivery, including topical administration. When a drug is delivered topically (e.g., patch or ointment), absorption is driven by the ability of the drug to cross the skin barrier and enter into circulation via capillaries in the skin. Often, this approach is used as a delivery method for extended release, in which a drug depot builds up under the skin and slowly makes its way to the blood vessels. Alternatively, respiratory administration of drugs is another approach for systemic absorption, as there is only a thin layer of cells lining the alveoli and a high density of capillary blood vessels. Respiratory administration can occur via vapors, gases, or aerosols. Other routes of administration that require absorption include SQ or IM injection, but the barriers to absorption are much less because injections typically are delivered into well-perfused tissue.

Bioavailability is an important measure of how much drug reaches the systemic circulation after absorption, which is particularly significant for oral, topical, and inhalation routes of administration. Only the bioavailable drug can exert its pharmacologic activity. This parameter is typically expressed as a percentage of drug relative to the dose; as intravascular administration is directed into the system circulation, its bioavailability is 100% by definition. For inhalation or topical administration, the bioavailability will be limited by the ability of the drug to exit the tissue depot into the systemic circulation and will be driven by the lipophilicity of the drug. For oral

administration, the bioavailability will be dependent not only on the absorption of the drug in the GI tract, but also on the first-pass metabolism prior to the drug reaching the systemic circulation. Bioavailability is determined by comparing the area under the time−concentration curve (AUC) for a particular route of administration to the AUC for an intravenous administration of the drug.

Distribution

Distribution encompasses the reversible transfer of drug between the compartments and includes delivery of a drug to its physiologic site of action. Drug distribution can occur between different compartments; in clinical pharmacology, compartments where drugs can partition include blood plasma, interstitial fluids, fat tissue, intracellular fluids, and transcellular (e.g., pleural and peritoneal) fluids. Blood plasma is the primary compartment utilized in pharmacokinetic modeling and is the basis of a one-compartment pharmacokinetic model. Several factors influence drug distribution, and the overall process is dependent on both the physiochemical properties of the drug as well as the compartments to which a drug will be deposited. Drug lipophilicity and polarity are primary considerations in distribution into various physiologic compartments. Other considerations that influence the rate and amount of drug delivered to a target organ or tissue include vascular permeability, perfusion of blood into various compartments, binding of the drug to plasma proteins or tissue components, and the permeability of tissues and cell membranes to the drug molecule. For instance, drugs are more easily delivered to highly perfused organs, including the heart, kidney, and hepatic system, as opposed to less well-perfused peripheral tissues (e.g., muscle and fat deposits).

Of particular interest for certain drugs is the binding to plasma proteins, particularly albumin and α1-acid glycoprotein. It is well known that the "free drug" fraction, or fraction of the drug that is not protein-bound, is the biologically active form of the drug that interacts with the drug target, and is bioavailable to be distributed across cell membranes and into tissues. Typically, the compartmentalization of unbound drugs occurs between the blood plasma and the target compartments until an equilibrium is achieved. For highly protein-bound drugs ($>90\%$ bound), changes in protein binding due to drug−drug interactions, posttranslational modifications, or underlying pathophysiologic states such as portal vein obstruction or malnutrition can significantly affect a drug's pharmacokinetic profile and, ultimately, the patient's response to the drug. Discussion of free drug monitoring for therapeutic drug management can be found in Chapter 51, Therapeutic drug monitoring.

Metabolism

Metabolism is the process of converting one chemical species into another chemical species; this is primarily enzyme-driven. Drug metabolism reactions in PK can be classified as Phase I or Phase II reactions. The cytochrome P450 (CYP 450) enzyme superfamily is the set of enzymes primarily responsible for drug metabolism in the human body and is the main source of genetic variability in PK. Drug metabolism may result in the deactivation of a drug by converting it into a biologically inactive derivative of the parent molecule. However, some metabolites retain a fraction of parent drug biologic activity, and some are equally potent. In those cases, both drug and metabolite may be measured for therapeutic management.

Conversely, many drugs can be administered as prodrugs and are activated in vivo via enzyme catalytic activity. In such cases, the active metabolite may be more informative in understanding downstream pharmacologic effects. Therefore it should not be assumed that in vivo metabolism leads to decreased drug activity or potency; for example, therapeutic compounds such as azathioprine and tenofovir disoproxil fumarate, which are used for treatment of certain childhood leukemias and HIV, respectively, are administered as prodrugs, and are biologically activated via specific enzymatic cascades. Notably, neither drug is modulated by the CYP 450 family, and bioactivation occurs external to the hepatic system. Thus a key component of clinical pharmacology is to understand the mechanistic pathways of administered drugs. Common Phase I and Phase II reactions are shown in Table 50.1.

Phase I reactions include oxidation, reduction, and hydrolysis, and hydration—oxidation reactions are the most common. Mechanistically, Phase I oxidation enzymes can catalyze a monooxygenase reaction, which involves the incorporation of oxygen into an organic substrate, and

TABLE 50.1 Phase I and Phase II metabolic reactions.

Phase 1 reactions	Phase 2 reactions
Oxidation	Glucuronidation
Reduction	Glucosidation
Hydrolysis	Sulfation
Hydration	Acetylation
Dethioacetylation	Methylation
Isomerization	Glutathione conjugation
	Amino acid conjugation
	Fatty acid conjugation

the reduction of a second molecule of oxygen to water. The enzymatic reactions for Phase I metabolism occur primarily in the endoplasmic reticulum of hepatic microsomes. The majority of Phase I metabolism reactions (~80%–90%) in humans are catalyzed by the CYP 450 family of enzymes. In many cases, the Phase I metabolic reaction is a precursor to the Phase II reaction; in other words, the Phase I product is the substrate for a Phase II reaction.

CYP 450 enzymes are hemeproteins that include a ferriprotoporphyrin-9 core and are named for their characteristic absorbance at 450 nm due to the presence of this core. This family of enzymes is classified by their sequence homology; CYP enzymes with >40% homology belong to the same family and those with >55% homology are considered part of the same subfamily. When naming the enzymes, a number will be included at the end to denote isoforms from the same gene. The naming convention for isoforms of CYP enzymes is illustrated in Fig. 50.1.

Common drug metabolizing enzymes include CYP2D6, CYP2C9, CYP2C19, and CYP3A4. There is significant genetic variability in CYP 450 enzymes, and these polymorphisms are often responsible for interindividual variability in drug metabolism (for further discussion, see Chapter 53: Pharmacogenomics). In addition, CYP 450 enzymes are particularly susceptible to enzyme induction and inhibition (discussed below), meaning that drugs metabolized by these enzymes carry significant risk of drug–drug interactions when multiple drugs are ingested, as enzymatic activity can be modulated via positive or negative feedback mechanisms.

Phase II reactions are conjugation reactions, which couple the drug or metabolite with endogenous compounds to increase water solubility and polarity. These reactions primarily include glucuronidation, sulfation, methylation, and acetylation, as well as conjugation of amino acids, fatty acids, or glutathione. Phase II conjugation reactions can occur directly to the drug molecule or, as noted previously, to a molecule generated by a Phase I metabolic reaction. Phase II reactions deactivate the drug, either by decreasing receptor affinity or by enhancement of excretion from the body. Several phase II enzymes have been associated with genetic polymorphism, including uridine diphosphate glucuronosyltransferase (UGT), *N*-acetyltransferase (NAT) as well as other acetyltransferases, glutathione S-transferases, and methylases. To date, only polymorphisms in UGT and NAT enzymes have demonstrated clinical significance.

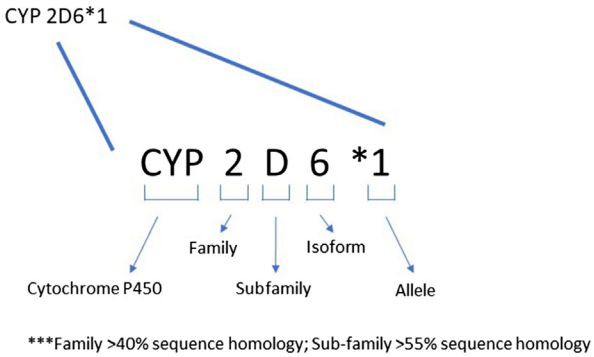


FIGURE 50.1 Cytochrome P450 enzyme naming convention. ***Family >40% sequence homology; subfamily >55% sequence homology.

Drug metabolism can be substantially influenced by enzyme induction or inhibition; many drug–drug interactions occur because of their inductive or inhibitory effects on drug metabolizing enzymes. In addition, food–drug interactions, herbal supplement interactions, and even lifestyle interactions (e.g., smoking and clozapine [4]) must be considered. Enzyme induction can occur through increased enzyme synthesis, protein stabilization, or activation of other components in the reaction—or some combination of these mechanisms. Enzyme inhibition can occur through competitive inhibition, conformational changes from binding to a nonactive site, accelerated protein degradation or destruction, or suppression of protein synthesis. There are several excellent resources where specific drug interactions are described in greater detail [5,6].

As previously mentioned, metabolism may also lead to the activation of a drug when the drug is administered in an inactive or prodrug form. This has been exploited as a drug delivery strategy and may be done to improve drug delivery or PK, to decrease toxicity, or to target the drugs to a specific type of tissue or cell in the body. Common examples include fosphenytoin, azathioprine, codeine, and primidone.

Excretion

Elimination and excretion are often used interchangeably; however, one is a subcategory of the other. Elimination refers to the irreversible removal of the drug from the compartment of measurement (primarily blood); this process can be achieved via metabolism (as previously discussed) and/or through excretion, which is the elimination of the drug or metabolite without chemical modification. The primary route of excretion is renal filtration; however, drugs or metabolites can also be eliminated through biliary excretion in the liver. In general, small, polar molecules (<200 Da) are excreted through the kidney, while larger or amphoteric substances are excreted via the liver. There can also be drug or metabolite excretion through the lungs, sweat, or mucosal secretion, but these are all minor pathways of clearing a drug from the body.

Drugs and metabolites are normally excreted by the kidney through a combination of three distinct processes: (1) glomerular filtration; (2) active tubular absorption/

secretion; and (3) passive, flow-dependent back diffusion. Glomerular filtration involves the unidirectional removal of a compound from blood and is predominantly dependent on renal perfusion and molecular size; only nonprotein bound drugs are filtered through the glomerulus. Renal secretion and absorption are energy-driven processes in the renal proximal tubules that rely on the activity of organic anion and cation transport receptors. Within the proximal tubules, small solutes, including small exogenous molecules, can be absorbed from the renal filtrate. Conversely, molecules like hydrogen ions and creatinine can be secreted into the proximal tubules, contributing to the renal filtrate. Passive back diffusion, or nonenergy-dependent reabsorption, also occurs in the renal tubules and is dependent on lipophilicity, filtrate pH, and urine flow rate. At low flow rates, there is a greater opportunity for diffusion of a drug from the renal filtrate back into circulation. The overall renal excretion of a drug (or metabolite) is the sum of filtration and secretion, minus reabsorption.

Given the importance of renal function in drug clearance; serum creatinine and creatinine clearance (see Chapter 35: Laboratory evaluation of kidney function) have been used as markers for drug clearance and safety. However, this was displaced by estimated creatinine clearance (eCrCl), originally calculated using the Cockcroft−Gault equation (C-G), which was developed in a small, homogenous population [7]. Currently, recommendations regarding drug dosing practices with respect to renal function are based on creatinine clearance determined by the C-G equation. This has led to challenges, as more accurate estimates of glomerular function have been determined and implemented as a standard of practice. Currently, most clinical laboratories provide estimated glomerular function using the modification of diet in renal disease [8] or chronic kidney disease-epidemiology collaboration (CKD-EPI) [9] equations; however, many dosing recommendations still refer back to the C-G estimations of renal function. Currently, the National Kidney Disease Education Program (NKDEP) has not endorsed any particular equation, as in many populations there is little difference in dosing based on the estimated glomerular filtration rate (eGFR) equation used. The NKDEP guidance recommends the use of either eGFR or eCrCl for drug dosing and to consider exogenous filtration markers, or a timed assessment of CrCl when dosing medications within narrow therapeutic indices or with high toxicity [9]. However, the National Kidney Foundation has supported the use of the CKD-EPI equation for clinical implementation due to its improved performance in estimating GFR.

In addition to protein synthetic and metabolic functions, the liver (see Chapter 31: Laboratory diagnosis of liver disease) also has an excretory function, where bile is produced and excreted. The physiologic functions of bile are as follows: (1) facilitate intestinal absorption of ingested lipids; (2) cholesterol elimination to maintain homeostasis; and (3) excretion of biotransformed waste products. In the context of drug excretion, some parent drugs and metabolites are excreted via bile into the intestinal tract. From there, they can be reabsorbed into the blood resulting via the enterohepatic recirculation cycle, or they can be excreted in stool. Enterohepatic recirculation can be significant for pharmacokinetic analysis, as it can result in a secondary peak for the drug if recirculation is significant.

As an extension of both distribution and clearance, drugs may be deposited in compartments that are ultimately excreted and transmitted to another individual. The chief example of this would be the partitioning of drugs into breast milk, resulting in neonate or infant exposure to a potentially active agent. The partitioning of a drug into breast milk is predicated on a number of factors, including drug distribution parameters such as drug lipophilicity and protein binding. From a matrix perspective, breast milk contains proteins, lipids, carbohydrates, and other nutrients; carbohydrate and lipid loads are higher in breast milk than blood. As such, drug delivery to breast milk will be dependent on penetration into a lipophilic compartment.

The distribution of drugs into breast milk can have significant consequences [10]. A case study in the mid-2000s described the neonatal toxicity and death due to the increased delivery of the codeine metabolite morphine into breast milk. In this case, the mother was prescribed acetaminophen-containing codeine for postpartum pain; she had a gene duplication for the CYP2D6 gene, which is associated with an ultrarapid metabolizer phenotype and increased metabolism of codeine to the potent morphine metabolite. The infant was exposed to lethal concentrations of morphine and expired shortly postbirth [11]. This case led to a black box warning on codeine, which directed limiting the use of codeine as an analgesic in postpartum mothers to the lowest dose for the shortest amount of time. In many settings, codeine is not prescribed to mothers for postpartum pain. In order to assist clinicians, within the NIH Toxicology Data Network, there is a lactation database that has important pharmacologic information on drug concentrations observed in breast milk and infant serum, as well as potential associated toxicities [12]. Understanding neonatal exposure to therapeutic (or illicit) substances is an important component of promoting infant health.

Pharmacokinetic calculations

For most calculations of pharmacokinetic parameters, the most straightforward model, a one-compartment model, is employed. While more complex models can be

constructed, this model is most commonly used for TDM and toxicology and treats the body as a single compartment, where the concentration of the drug in the blood tracks with a proportional change of concentration in all tissues of the body. Two-compartment models can also be used, where the first compartment is designated to be in rapid equilibrium with the blood, and a second compartment (e.g., adipose tissue) is present with a much slower equilibrium between it and the first compartment. A comparison of one- and two-compartment models is illustrated in Fig. 50.2. Similarly, while complex mathematical models can be employed, the kinetics are most often assumed to be zero-order or first-order kinetics. In zero-order kinetics, the rate of excretion or elimination is fixed, and independent of the amount of substance available. In first-order kinetics (also known as linear kinetics), there is a proportional relationship between the absorption and elimination rates and the amount of substance present; the rate constant that characterizes this proportionality represents a number of physiologic parameters. Most therapeutic substances of clinical interest demonstrate first-order kinetics, and common pharmacokinetic calculations describe these situations. It is important to note that for many drugs, mixed or capacity-limited kinetics are observed in the context of saturable systems, such as when the concentration of a drug or intermediate is present in higher concentrations than the enzymatic capacity for metabolism. In such cases, pharmacokinetic calculations are applied assuming that the system is in the linear range of its kinetics. Graphical examples of zero-, first-, and mixed-order kinetics are illustrated in Fig. 50.3.

Based on these simple physiologic and mathematical models, there are a few basic pharmacokinetic calculations that every clinical chemist should be able to perform:

Elimination constant

The elimination constant (k_{el}) is a simple calculation determined by measuring the concentration of drug in the blood at two time points during the terminal elimination phase:

$$k_{el} = \frac{\ln(C_1/C_2)}{t_2 - t_1}$$

where t_1 is the first time point, t_2 is the second time point, C_1 is the concentration measured at t_1, and C_2 is the concentration measured at t_2.

Elimination half-life

The elimination half-life is the amount of time required for 50% of the drug to be removed from the blood during the terminal elimination phase. This can be determined graphically from a linear plot of time versus log

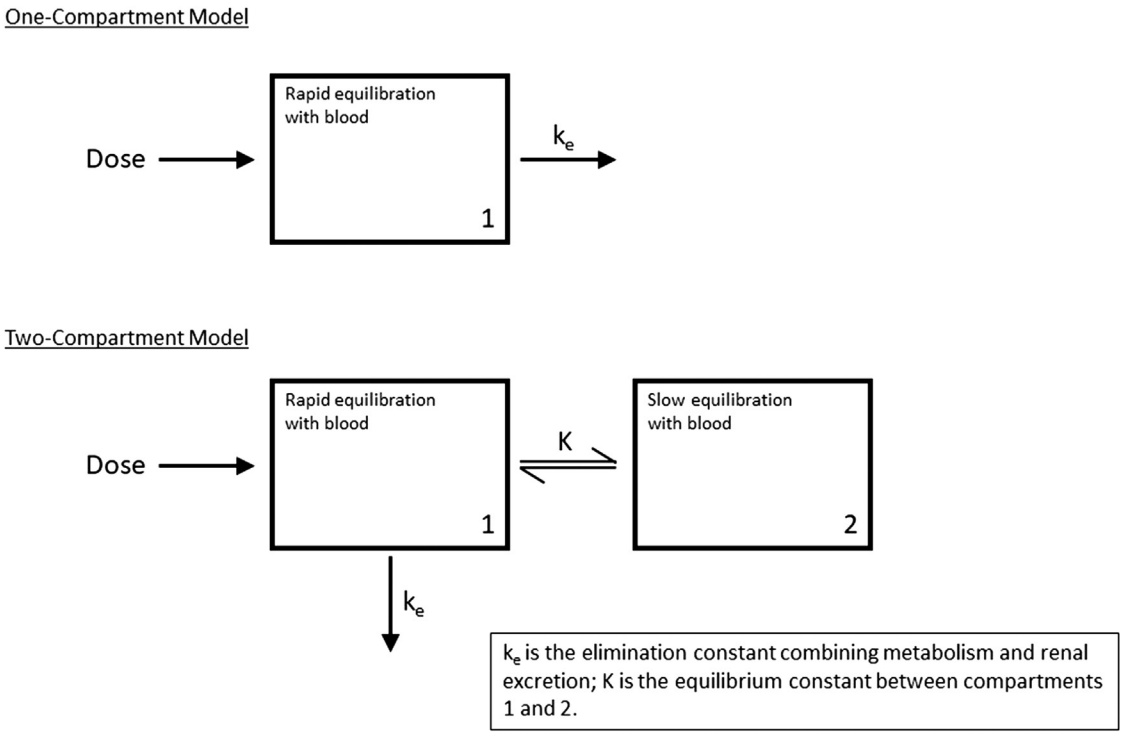


FIGURE 50.2 One- and two-compartment model schematic. k_e is the elimination constant combining metabolism and renal excretion and K is the equilibrium constant between compartment 1 and 2.

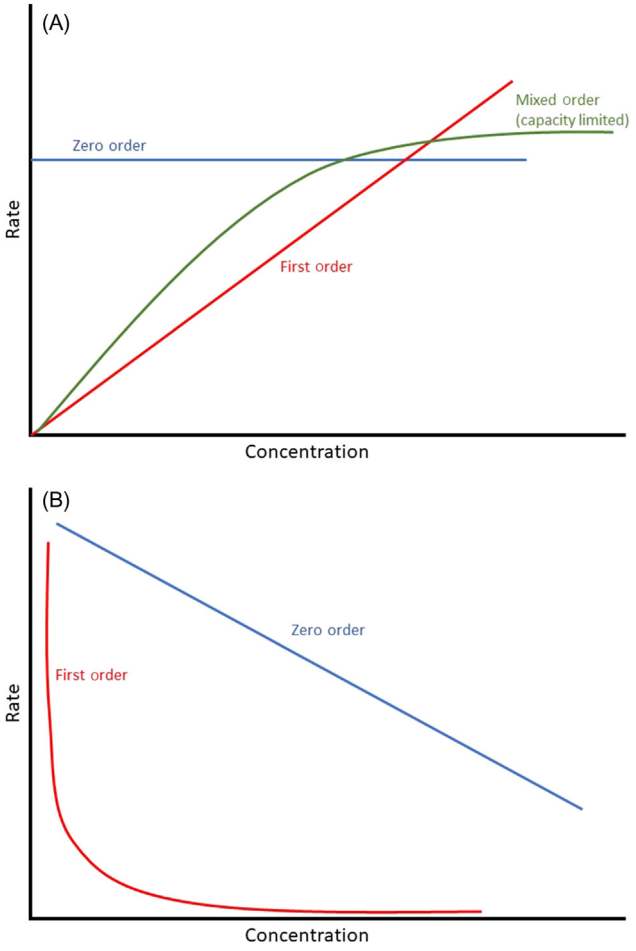


FIGURE 50.3 (a) Zero-, first-, and mixed-order absorption kinetics. (b) Zero- and first-order elimination kinetics.

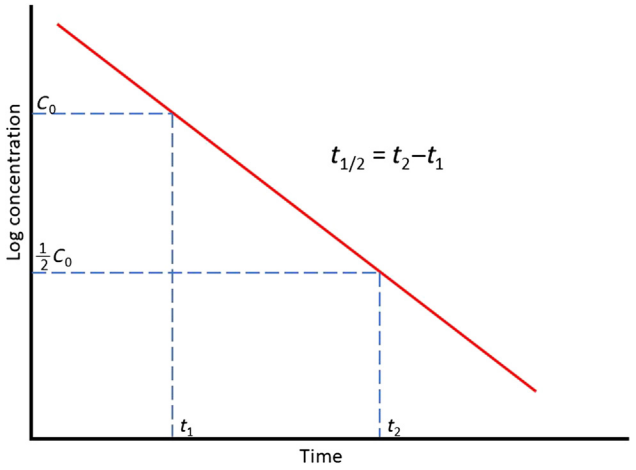


FIGURE 50.4 Graphical determination of half-life ($t_{1/2}$).

concentration (Fig. 50.4), or it can be calculated using the equation below:

$$t_{1/2} = \frac{0.693}{k_{el}}$$

In addition, based on the elimination half-life, the time needed for clearance of a drug can be estimated; approximately, 97% of the drug is eliminated by five half-lives, and just over 99% of the drug is eliminated when seven half-lives have passed.

Volume of distribution

The volume of distribution (V_d) is a descriptive parameter that relates the amount of drug distributed into tissue and intracellular fluid to the amount that remains in the blood compartment. V_d is expressed as the amount of fluid that would hypothetically be necessary to contain all of the drug administered. It is calculated using the equation below:

$$V_d = \frac{\text{Dose}}{C_{ss}}$$

The V_d can be compared with fluids in the body to get an idea of where a drug is distributed. The plasma volume is approximately 0.4 L/kg, and the total body water is approximately 0.6 L/kg; drugs that are primarily distributed in tissue will have V_d greater than these amounts.

The V_d for many drugs are published and available in the scientific literature, but it is important to note that these are population estimates only and have been typically established in healthy populations. Any deviation from the "normal" body composition due to population differences (e.g., neonates and pregnancy) or pathophysiology (e.g., burns, edema, pleural effusion, or ascites) will result in a V_d deviation from the published value. V_d can be used to estimate the starting dose necessary to achieve a target blood concentration by rearranging the equation above, and can also be used to estimate the replacement drug needed after dialysis.

Clearance

The clearance reflects excretion of the drug from the body by all routes and is linear (first-order) for most drugs. Clearance represents the amount of blood cleared of a drug per unit of time, and it is expressed in L/kg/h or mL/min. Clearance can be calculated in a number of ways, including:

$$\text{CL} = \frac{\text{Dose/dosing interval}}{C_{ss}};\ \text{or CL} = k_{el} x V_d;\ \text{or}$$
$$\text{CL} = \left(0.693/t_{1/2}\right) \times V_d$$

where C_{ss} is the steady-state concentration of the drug, k_{el} is the elimination constant, V_d is the volume of distribution, and $t_{1/2}$ is the elimination half-life.

Steady-state dose adjustment

For most drugs, the desired clinical response is not achieved with a single dose, so multiple doses are required. With multiple doses, the drug accumulates and increases in blood concentration until the amount of drug given is equal to the amount that is cleared; this is known as steady state. When the drug is administered in intervals that approximate the elimination half-life of the drug, steady state is typically achieved in seven doses (or seven half-lives); although for practical intents, five doses or half-lives are sufficient to assume steady state has been achieved.

Once steady state has been achieved, applied PK (or TDM, Chapter 51: Therapeutic drug monitoring) can be used to adjust the drug dosage to achieve a desired target concentration in the blood, provided the drug follows first-order kinetics. Calculation of the new dose can be performed using the equations below:

$$\text{Current dose}/\text{measured blood concentration} = \text{New dose}/\text{desired concentration}$$

so

$$\text{New dose} = \left(\text{Desired concentration}/\text{measured blood concentration}\right) \times \text{Current dose}$$

It is important to note that this approach cannot be used if the drug follows zero-order kinetics or if the patient has not achieved steady state. If the patient is not at steady state but the drug follows first-order kinetics, it is important to wait until the fifth dose is administered to utilize TDM and adjust the dose. When zero-order or nonlinear kinetics are present, the dose adjustment based on blood concentrations becomes an iterative, empiric process as the above proportionality equation does not apply. This does not invalidate the usefulness of dosing based on blood concentrations for nonlinear kinetics, as evidenced by common TDM applications such as phenytoin or methotrexate (see Chapter 51: Therapeutic drug monitoring).

Population pharmacokinetics and modeling

Although the aforementioned equations provide important information on pharmacokinetic parameters, more complex calculations and models are needed to characterize comprehensively drug PK, as there is significant variability in estimated pharmacokinetic parameters due to intra- and interindividual variabilities. This approach to PK has led to the specialized field of pharmacometrics, which is a multidisciplinary field that incorporates mathematical modeling, demographic information, drug concentrations, and patient health to estimate pharmacokinetic parameters on a population level, characterize pharmacokinetic-pharmacodynamic relationships, and generate mechanistic drug models. Population pharmacokinetics (PopPK) combines data (including laboratory, demographic, and pharmacy dosage information) from a large number of individuals into a single mathematical model. Physiologically based PK modeling (PBPK) utilizes complex models that account for specific anatomic and physiologic variables in addition to standard pharmacokinetic parameters. PBPK models employ extrinsic data (i.e., environmental factors, smoking, drug—-drug interactions, and patient demographics) to predict drug PK /ADME. Although this modeling approach is not used extensively in clinical laboratory medicine, PBPK models are common in pharmaceutical research and drug development. However, we can utilize the findings from these models to identify target concentrations for therapeutic compounds. Specialized software tools are typically required for the generation of popPK and PBPK models.

A type of model that has been used in the clinical laboratory setting is a Bayesian model. In Bayesian modeling, the pharmacometrician takes advantage of both the prior population-derived PK information as well as sparse data obtained from an individual patient or research subject. This is an iterative process where the initial therapeutic regimen is set using PopPK and then modified using new data from just a few patient blood samples.

Special populations

While most pharmacokinetic data are collected in healthy, early adult populations, it is important to note that these populations may not be the clinically relevant populations. While pharmacokinetic information can be applied to many patient populations, there are some special populations for which general pharmacokinetic models and assumptions for healthy adults do not apply. In addition to the populations described below, PK have been influenced by sex, ethnicity, and hormonal concentrations.

Neonate/pediatric

For pediatric patients and neonates in particular, there are significant differences in body composition and physiology that affect a drug's pharmacokinetic profile. In these patients, the drug dosage and dosing interval cannot be simply downadjusted by accounting for body mass (mg/kg). In general, pediatric patients have a lower amount of HCl in the stomach, slower gastric emptying, and lower production of bile acids, all of which significantly affect the absorption phase of PK relative to an adult population. In addition, pediatric patients typically have a higher percentage of body water and lower body fat, which affects

the distribution of drugs after absorption. The metabolic activity and adrenal function in neonates and children are significantly different from adults and can affect the metabolism phase of PK. For neonates with jaundice, the PK of highly protein-bound drugs (e.g., phenytoin) can be affected as bilirubin can compete with these drugs for binding sites on albumin—in these cases, free drug measurements are particularly important.

Geriatric

For older patient populations (>70 years of age), there are also significant differences relative to health. In patients with advanced age, there is decreased saliva and gastric HCl production, decreased gut motility, and variance in gastric emptying, all of which can significantly impact drug absorption. Older patients also tend to have increased body fat, decreased muscle mass, and decreased perfusion of tissues, all of which will alter drug distribution relative to healthy adults; older patients may also have a lower V_d than an adult population. With respect to drug metabolism, in geriatric patients, Phase I metabolism (e.g., CYP 450) is decreased, while Phase II metabolism (e.g., glucuronidation) is unchanged. Finally, renal function typically decreases with age, which affects the excretion of drugs and metabolites.

In addition to these physiologic factors, there are some important external factors to consider for this population. Polypharmacy is more common in geriatric patients, which can lead to increased risk of drug−drug interactions, and these patients are often taking nutritional supplements or complementary and alternative medicines that also can cause unintended drug interactions [1]. Older patients are also likely to be managed by multiple physicians and are more prone to self-medication or noncompliance with medications, all of which can complicate pharmacokinetic analysis. Given the prevalence of multiple pathologic conditions, the symptoms of drug toxicity can be similar to chronic disease symptoms or signs of aging, making the clinical assessment of drug response complex.

Obese

Obesity is defined as anyone with a body mass index greater than 30 kg/m^2, with serious obesity defined as greater than 40 kg/m^2. The prevalence of obesity and comorbidities associated with this condition is increasing globally and has significant implications for health care and include its impact on PK. While standard dosing of drugs is often achieved using body weight (mg/kg), it is generally recognized that at some point, there is a nonlinear relationship between the total body weight and the pharmacokinetic parameters. There are important changes in cardiac output, plasma protein concentration, and organ size and function that are associated with obesity, which will affect PK. A higher proportion of adipose tissue in this patient population leads to significant changes in drug distribution; the V_d for an obese patient may differ significantly from that of a nonobese person. Fatty liver disease can result in liver damage that could negatively impact metabolism. It should be noted that while total body weight does not have a linear relationship in obese patients, it has been demonstrated that lean body weight appears to maintain some degree of linearity with pharmacokinetic parameters and can be used in obese patients to optimize drug dosing [13]. The lean body weight can be calculated using the following equations [14]:

$$\text{LBW(male)} = \frac{9270 \times \text{TBW}}{6680 + (216 \times \text{BMI})}$$

or

$$\text{LBW(female)} = \frac{9270 \times \text{TBW}}{8780 + (244 \times \text{BMI})}$$

where LBW is Lean Body Weight, TBW is Total Body Weight, and BMI is Body Mass Index. The constants in this equation were derived empirically during development of the model.

References

[1] G.N. Asher, A.H. Corbett, R.L. Hawke, Common herbal dietary supplement-drug interactions, Am. Fam. Phys. 96 (2) (2017) 101−107.

[2] S.A. Abdulrahman, K. Ganasegeran, L. Rampal, O.F. Martins, HIV treatment adherence - a shared burden for patients, health-care providers, and other stakeholders, AIDS Rev. 21 (1) (2019) 28−39.

[3] L. Bouwman, C.M. Eeltink, O. Visser, J. Janssen, J.M. Maaskant, Prevalence and associated factors of medication non-adherence in hematological-oncological patients in their home situation, BMC Cancer. 17 (1) (2017) 739.

[4] M. Mayerova, L. Ustohal, J. Jarkovsky, J. Pivnicka, T. Kasparek, E. Ceskova, Influence of dose, gender, and cigarette smoking on clozapine plasma concentrations, Neuropsychiatr. Dis. Treat. 14 (2018) 1535−1543.

[5] S. Cole, E. Kerwash, A. Andersson, A summary of the current drug interaction guidance from the European Medicines Agency and considerations of future updates, Drug. Metab. Pharmacokinet. (2020).

[6] B. Zhou, R. Wang, P. Wu, D.X. Kong, Drug repurposing based on drug-drug interaction, Chem. Biol. Drug. Des. 85 (2) (2015) 137−144.

[7] D.W. Cockcroft, M.H. Gault, Prediction of creatinine clearance from serum creatinine, Nephron. 16 (1) (1976) 31−41.

[8] L.A. Stevens, T.D. Nolin, M.M. Richardson, H.I. Feldman, J.B. Lewis, R. Rodby, et al., Comparison of drug dosing recommendations based on measured GFR and kidney function estimating equations, Am. J. Kidney Dis. 54 (1) (2009) 33−42.

[9] L.A. Stevens, A.S. Levey, Measured GFR as a confirmatory test for estimated GFR, J. Am. Soc. Nephrol. 20 (11) (2009) 2305–2313.
[10] D.N. Bateman, M. Eddleston, E. Sandilands, Codeine and breast-feeding, Lancet. 372 (9639) (2008) 625.
[11] G. Koren, J. Cairns, D. Chitayat, A. Gaedigk, S.J. Leeder, Pharmacogenetics of morphine poisoning in a breastfed neonate of a codeine-prescribed mother, Lancet. 368 (9536) (2006) 704.
[12] Drugs and Lactation Database (LactMed) [Internet], Bethesda (MED): National Library of Medicine (US), 2006. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK501922/>.
[13] R. Roubenoff, J.J. Kehayias, The meaning and measurement of lean body mass, Nutr. Rev. 49 (6) (1991) 163–175.
[14] S. Janmahasatian, S.B. Duffull, S. Ash, L.C. Ward, N.M. Byrne, B. Green, Quantification of lean bodyweight, Clin. Pharmacokinet. 44 (10) (2005) 1051–1065.

Self-assessment questions

1. Which of the following types of drug formulation would be considered an intravascular route of administration?
   a. Extended-release capsule
   b. Transdermal patch
   c. PICC line
   d. Sublingual film
2. Phase 1 metabolic reactions include which of the following?
   a. Acetylation
   b. Hydrolysis
   c. Glucuronidation
   d. Methylation
3. A plasma concentration for an experimental drug is reported as 45 mg/L following a dose of 150 mg. The patient in this study is a 45-year-old male weighing 85 kg. Where is most of this drug found?
   a. Adipose tissue
   b. Plasma
   c. Urine
   d. Bile
4. Drug *X* is a highly protein-bound drug, primarily bound (>95%) to albumin. Which of the following will result in an increase in the free drug concentration?
   a. Hemolysis
   b. Hyperlipidemia
   c. Hypoglycemia
   d. Hyperbilirubinemia
5. Pediatric patients exhibit significant differences in PK compared with adults due to all of the following except:
   a. Cardiac output
   b. High proportion of body water
   c. Slower gastric emptying
   d. Metabolic differences
6. The primary route of elimination for most drugs is:
   a. Exhalation
   b. Biliary
   c. Renal
   d. Mucosal secretion
7. A patient has been administered a new drug, and the concentration in blood was measured at two time points during terminal elimination. At 4-hours postdose, the concentration was determined as 120 mg/L, and at 8-hours postdose, the concentration was determined as 55 mg/L. If the drug is discontinued, approximately how long will it take for 97% of the drug to be cleared?
   a. 17–20 hours
   b. 5–10 hours
   c. 2–4 hours
   d. 20–25 hours

Answers

1. c. PICC line. A PICC line is a peripherally inserted central catheter, that is, a venous catheter used for long-term IV drug or nutrition administration.
2. b. Hydrolysis. Phase I reactions include oxidation, reduction, and hydrolysis, while Phase II reactions are conjugation reactions such as the choices in A, C, and D.
3. b. Plasma. The calculated V_d = (dose/concentration)/patient weight. Dose = 150 mg; concentration = 45 mg/L, and patient weight = 85 kg. V_d = 0.039 L/kg, which is comparable with the average plasma volume (0.04 L/kg).
4. d. Bilirubin can compete with highly protein-bound drugs for albumin-binding sites.
5. a. Cardiac output. While cardiac output likely does vary between the kids and the adults, there can also be significant variability within the adult population, so cardiac output is not likely the reason for these population difference compared with the other three noted differences.
6. c. Renal. Most drugs are excreted renally either unchanged or metabolized followed by renal elimination of the metabolites.
7. a. 17–20 hours. First, the k_{el} is calculated from the two concentration measurements, and then, the half-life is calculated:

$$k_{el} = \left(\ln\left(C_1/C_2\right)\right)/(t_2 - t_1)$$
$$= \left(\ln\left(120/55\right)\right)/[8 - 4\text{hours}] = \ln(2.18)/4 = 0.19$$

$$t_{1/2} = 0.693/k_{el} = 0.693/0.19 = 3.65\text{hours}$$

97% of the drug is cleared in 5 half-lives, so it will take approximately 18 hours.

Chapter 51

Therapeutic drug monitoring

William Clarke

Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD, United States

Learning objectives

After reviewing this chapter, the reader should be able to:

- Define therapeutic drug monitoring (TDM).
- Describe the challenges of managing drugs with significant pharmacokinetic variability.
- List drugs that are commonly managed using TDM.
- Discuss the analytical methods available for TDM.

Management of therapeutic drugs

The vast majority of drugs are managed in a very standard way—the drugs are dosed on a unit per body mass (e.g., mg/kg) basis, and then, the dosage is adjusted based on the clinical response as empirically assessed by a physician. This is often described as "titration to clinical effect." However, for a small subset of drugs, the clinical effects (pharmacodynamics) can be assessed objectively through laboratory measurements—for instance, warfarin and assessment of coagulation via international normalized ratio measurement, or statin drugs and blood lipid levels. For an even smaller subset of drugs, the pharmacokinetic−pharmacodynamic relationship is not predictable from the dose, and the pharmacokinetics are highly variable between the individuals—for these drugs, management can be particularly challenging. It has been recognized that for drugs with high pharmacokinetic variability, measurement of serum concentrations to assess the patient's individual pharmacokinetics (see Chapter 49, The porphyrias: fundamentals and laboratory assessment) relative to a population-based pharmacokinetic profile is useful to avoid possible toxic side effects or to ensure that patients have achieved a stable therapeutic level. This use of applied pharmacokinetics is primarily referred to as therapeutic drug monitoring (TDM).

Principles of therapeutic drug monitoring

TDM in practice is performed by collection of a blood sample at a known time relative to administration of the last (or next) dose. The concentration of drug and/or metabolite is measured in the sample and compared with a target range or the predicted pharmacokinetics for the drug. In order for TDM to be effective and necessary, several criteria must be met: (1) the drugs must have a narrow therapeutic index (the difference between the minimum effective concentration and the toxic concentration, relative to the pharmacokinetic variability); (2) the relationship between the drug dose and the blood concentration must be highly variable and/or not predictable interindividual variability; (3) the relationship between blood concentration and clinical or toxic effect must be well defined; (4) there should be serious consequences for under- or overdosing; and (5) the result of TDM testing must be interpretable and actionable—there should be an effect on clinical outcomes. TDM results should answer a clinical question. In addition, TDM can be useful if the drug in question is administered with other compounds that potentially interact and affect the pharmacokinetics. A list of commonly monitored drugs can be found in Table 51.1.

There are some drugs that meet most of the criteria for TDM, but for which measurements of drug concentrations are not particularly useful. For instance, when the drug is administered as a prodrug—the biologically active form of the drug is a metabolite of the administered substance—measurement of the administered drug does not help with guidance of therapy and the active metabolite must be measured (e.g., irinotecan and SN-38). In other cases, the drug is converted into its active form intracellularly, so systemic blood measurements in serum or plasma are not necessarily reflective of the therapeutic activity (e.g., nucleotide/nucleoside reverse transcriptase inhibitors and peripheral blood mononuclear cells). In addition, drugs for which tolerance can be developed (e.g., narcotics and pain management) prevent the utility of TDM, as the effective blood concentration range is a moving target and not stable within an individual patient.

While the primary function of TDM is to allow for adjustment of the drug dose, there are other applications

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00051-X

TABLE 51.1 Therapeutic ranges for commonly monitored drugs.

Drug class	Drugs	Therapeutic range	Notes
Antiepileptics	Phenytoin	10–20 μg/mL	CNS toxicity >30 μg/mL
	Carbamazepine	4–12 μg/mL	
	Primidone	5–12 μg/mL	Phenobarbital is an active metabolite and should be monitored too
	Phenobarbital	15–40 μg/mL	Significant CNS depression >40 μg/mL
	Valproic acid	50–100 μg/mL	
	Lamotrigine	2.5–15 μg/mL	
	Levetiracetam	12–46 μg/mL	
	Gabapentin	2–20 μg/mL	
	Oxcarbazepine	3–35 μg/mL	Therapeutic range is for the active metabolite, MHD
	Topiramate	5–20 μg/mL	
Immunosuppressants	Tacrolimus	5–15 ng/mL	
	Cyclosporine A	100–400 ng/mL	
	Sirolimus	4–12 ng/mL	
	Everolimus	3–15 ng/mL	
	Mycophenolic acid	AUC = 30–60 mg/h.L	Trough = 1–4 μg/mL
Cardioactive drugs	Procainamide	4–8 μg/mL	NAPA = 10–20 μg/mL; 5–30 μg/mL combined
	Lidocaine	1.5–5 μg/mL	
	Quinidine	2–5 μg/mL	
	Digoxin	0.8–2 ng/mL	
Psychoactive drugs	Amitriptyline + metabolite	80–200 ng/mL	Active metabolite = nortriptyline
	Nortriyptline	70–170 ng/mL	
	Imipramine + metabolite	175–300 ng/mL	Active metabolite = desipramine
	Desipramine	100–300 ng/mL	
	Lithium	0.4–1.2 mmol/L	
Antimicrobial drugs	Gentamicin	<2 μg/mL, trough	5–10 μg/mL, peak
	Tobramycin	<2 μg/mL, trough	5–10 μg/mL, peak
	Amikacin	<5 μg/mL, trough	20–25 μg/mL, peak
	Vancomycin	10–20 μg/mL	
	Voriconazole	1–5.5 μg/mL	
Chemotherapy drugs	Methotrexate	5–10 μM, 24-h postdose	0.5–1 μM, 48-h postdose; 0.05–0.1 μM, 72-h postdose
	Busulfan	900–1500 μmol.min	
Bronchodilators	Theophylline	10–20 μg/mL	
	Caffeine	8–20 μg/mL	

CNS, Central nervous system; *MHD*, monohydroxycarbamazepine; *NAPA*, N-acetylprocainamide.

of TDM results (e.g., questions to be answered). For example, TDM in antiretroviral management in patients with HIV is not really necessary for optimization of dose; however, it is vitally important to confirm adherence to the prescribed drug regimen in the context of increasing viral loads and apparent therapeutic failure. The TDM results can be used to assist in the determination of whether the patient has developed viral resistance to the prescribed drugs or whether they have just stopped taking their drugs. Another question that can be answered is whether drug is being absorbed at all. For instance, gastrointestinal inflammation could prevent drugs administered orally from entering the circulation. It is important to remember that while TDM is a tool for assessing the clinical presentation of the patient, it is not intended to be interpreted in a vacuum. Most therapeutic target intervals were not derived from large clinical studies, but are based on the observations and data from a single site. The absence of a therapeutic interval does not mean that TDM lacks utility, provided that there are specific criteria for the interpretation of the result and it will aid in clinical decision-making.

Analytical and clinical laboratory considerations for therapeutic drug monitoring

Much of the early work in TDM was performed using high-performance liquid chromatography or gas chromatograpy techniques (Chapter 8, Chromatography and electrophoresis); however, TDM became much more accessible as commercial immunoassays (Chapter 12, Immunoassays) were developed by a number of vendors. Unfortunately, development of commercial assays for TDM has stagnated, whether due to the relative small size of potential markets or perception that the role of pharmacokinetics in personalized medicine will decrease based on advances in genomics. Based on this, increased research effort in the area of assay development using noncommercial platforms is rapidly growing, especially in the area of liquid chromatography–tandem mass spectrometry (LC-MS/MS). While immunoassays provide rapid results and can be automated using instrumentation readily available in the clinical laboratory, chromatography-based methods (especially LC-MS/MS) often provide greater sensitivity and specificity, as well as the ability for rapid development of assays in response to clinical needs—although often at the cost of throughput and ease of use. A modern clinical laboratory performing TDM uses a combination of these technologies to provide services, with reference laboratories becoming more and more important for applications that are infrequently utilized in a single hospital or medical care setting.

Typically, TDM is performed using serum or plasma samples collected once the patient has reached steady state (see Chapter 50, Pharmacokinetics). An example of a therapeutic range relative to steady-state concentrations can be seen in Fig. 51.1. Steady state is achieved after seven half-lives of the drug has been administered to the patient; however, in practice, five half-lives is an acceptable surrogate (97% of steady state is achieved) to estimate a steady-state concentration. In some special cases, drugs are distributed primarily in the cellular component of blood (e.g., immunosuppressive drugs into red blood cells), and in those cases, anticoagulated whole blood is the preferred specimen. For TDM measurements, serum or plasma separator tubes are typically avoided, as the gel in these tubes may nonspecifically bind drugs in the sample and alter the concentration. The laboratory should also ensure they are aware of potential interferences from tube additives (e.g., citrate and phenytoin) and that appropriate containers are used during specimen collection. Once the patient has reached steady state, the

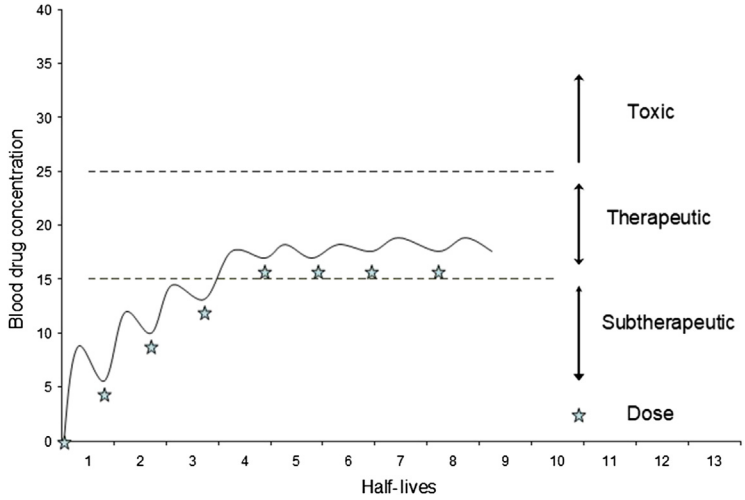


FIGURE 51.1 Illustration of steady-state drug concentrations as a function of drug half-life.

blood sample for TDM is typically drawn for a predose assessment (often referred to as a "trough" concentration). This type of sample is preferred, because measurements of peak concentrations are problematic and highly variable due to variability of other steps in the process—primarily variability in dose administration and specimen collection, with inconsistent documentation. The predose specimens are generally collected 30−60 minutes prior to the next scheduled dose, where the pharmacokinetics are the least variable. In some cases, peak or postdose specimens are analyzed, but this is less common and specific examples will be discussed later in the chapter.

For interpretation of the TDM results, it is vital that the time of the last dose and the time of specimen collection are captured as accurately as possible, regardless of whether a peak or predose measurement is used. While generally applicable therapeutic ranges (or target ranges) are provided with the result, it is important to note that this is a population-specific guidance and not a patient-specific guidance; the therapeutic target for a specific patient can vary based on their disease, body composition, formulation of the drug given to the patient, and many other factors (see Table 51.2). The optimal concentration for the patient is ultimately determined by the person providing care for that patient. In some instances, for highly protein-bound drugs, the free drug concentration is measured rather than the total drug concentration, as that is the biologically active form of the drug. In these cases, the sample is typically prepared using ultrafiltration to remove proteins (including protein-bound drug) from the sample, and free drug is subsequently measured in the ultrafiltrate. In addition, for some drugs, it is more appropriate to measure the active metabolite rather the parent drug—whether that parent drug is active or whether the parent drug is a prodrug as part of the drug formulation and delivery strategy. When both the parent drug and metabolite have biologic activity, both should be measured and a therapeutic range incorporating both should be established.

TABLE 51.2 Factors that influence TDM results.

Drug	Formulation
	Route of administration
	Dose regimen
	Pharmacokinetics: Vd, half-life, metabolites, etc.
Patient	Age
	Body composition
	Renal function
	Hepatic function
	Compliance
	Pregnancy
	Protein status
	Inherited variability in drug handling processes (pharmacogenetics)
	Disease
Specimen	Collection tube material, preservatives, and separator gels
	Time collected relative to dosing
	Sampling methods
	Storage
	Handling
Analytical method	Preanalytical processing (e.g., filtration and extraction)
	Densitivity
	Specificity
	Matrix effects
Other	Concomitant medications
	Supplements
	Diet
	Clerical errors

Clinical areas where therapeutic drug monitoring is routine practice

Epilepsy

There are many different drugs available for treatment of seizures—these drugs typically work by acting on neurotransmission pathways in the brain. In the treatment of epilepsy, TDM has a significant value, because the drugs act upon the central nervous system (CNS) to treat seizures, but they also can cause toxic CNS effects. In some cases, the drugs can cause seizures, making it difficult to differentiate subtherapeutic effects from toxic effects. TDM is useful to determine quickly whether drug concentrations are too low or too high upon initiation of therapy. In addition, once a baseline concentration has been established for patients on chronic therapy, blood levels can be used to investigate loss of seizure control or unexpected toxicity in a stable patient.

Phenytoin and *fosphenytoin* (prodrug for phenytoin) are widely used for treating epilepsy in a broad range of populations. Phenytoin has been in use for epilepsy since 1936; fosphenytoin was developed in 1996. Phenytoin is used primarily for tonic−clonic seizures and is not effective for absence seizures. The drug has a high bioavailability, but the rate of absorption is highly dependent on formulation—in some cases, absorption can occur up to

48 hours after administration. There is significant pharmacokinetic variability as phenytoin is metabolized by CYP2C19 and 2C9, is susceptible to drug−drug interactions, and is highly (approximately 90%) protein-bound. The typical target interval for this drug is 10−20 μg/mL, with CNS toxicity typically observed with concentrations >30 μg/mL [1]. Phenytoin is highly protein-bound, and if the clinical presentation does not match the total drug concentration, the free drug concentration is often measured in addition to the total. The target concentration for free (unbound) phenytoin is generally listed as 1−2.5 μg/mL.

Carbamazepine is a drug that was introduced for treatment of epilepsy in 1962. It is used in the treatment of generalized seizures, as well as partial and partial-complex seizures. This drug is CYP3A4-metabolized and affected both by inhibitor and inducers of this enzyme family—interestingly, its inductive capabilities lead to a drug−drug interaction with itself for patients on chronic therapy, giving increased blood concentrations. The therapeutic interval is defined as 4−12 μg/mL [2], and blood concentrations greater than 12 μg/mL are associated with adverse CNS effects, including respiratory depression, convulsions, and coma.

Primidone is a barbiturate drug used primarily for treatment of epilepsy in children; it is also used in the treatment of essential tremor. Primidone is primarily metabolized in the liver to phenobarbital via CYP2C19, and given that both parent drug and primary metabolite are biologically active, both must be monitored when primidone is used for antiepileptic therapy. The therapeutic target for primidone is 5−12 μg/mL and adverse effects are seen at concentrations greater than 15 μg/mL [3]. Clinical toxicity is primarily due to accumulation of phenobarbital and is manifested as reduced respiratory function and CNS depression. Primidone is CYP450-metabolized, and it is susceptible to drug−drug interactions.

Phenobarbital is another barbiturate drug used for treatment of epilepsy; as noted above, it is also the primary biologically active metabolite of primidone. Phenobarbital can be used alone as a therapeutic compound or in conjunction with other antiseizure medications. It is a broad-spectrum antiepileptic drug used to manage all seizure types with the exception of absence seizures. Phenobarbital is not extensively protein-bound (40%−60% bound to albumin); therefore measurement of free drug concentrations is not required to assess therapeutic efficacy. The therapeutic interval for phenobarbital is 15−40 μg/mL, with blood concentrations of greater than 40 μg/mL, leading to significant CNS depression [4]. Phenobarbital has a significantly longer half-life than many drugs that are monitored and is thus not as sensitive to timing of collection as other drugs.

Valproic Acid is a drug that has long been used for treatment of epilepsy (since 1962; FDA cleared in 1978), but is also used for the treatment of bipolar disorder, migraines, and neuropathic pain. In epilepsy, it us used primarily for treatment of absence seizures, but has also been used for the treatment of generalized tonic−clonic and complex partial seizures. As with phenytoin, valproic acid is highly protein-bound and measurement of the free (or unbound) drug can be used for patient management. The therapeutic interval is defined as 50−100 μg/mL, with concentrations greater than 100 μg/mL associated with significant toxic effects including hepatic, gastrointestinal, and CNS toxicity [5].

Lamotrigine is one of the newer generations of antiepileptic drugs; it was introduced for treatment of epilepsy in the early 1990s. The mechanism of action for lamotrigine is inhibition of glutamate and aspartate release. Currently, it is not only used for its antiseizure effects, but it is also used as a mood stabilizer in treatment of bipolar disorder and clinical depression. It is a broad-spectrum antiepileptic drug used for all types of seizures. The therapeutic interval is generally reported as 2.5−15.0 μg/mL (C_0 concentration) [6]. Lamotrigine also demonstrates enzyme autoinduction effects, with up to 20% reduction of blood concentrations occurring after 2 weeks of therapy. Adverse effects from elevated lamotrigine concentrations include vision abnormalities and GI toxicity.

Levetiracetam is another newer generation antiepileptic drug used for treatment of partial myoclonic, tonic−-clinic, and partial seizures. It is also used for the treatment of manic states in bipolar disorder and for migraine headaches. Levetiracetam has a relatively short half-life (∼6−7 hours) and is cleared renally, so any renal dysfunction or acute kidney injury may warrant TDM and/or dose adjustment. The generally reported therapeutic interval for C_0 samples is 12−46 μg/mL; no toxic threshold has been established [7]. The minimum concentration associated with seizure control is 3 μg/mL. Clinical toxicities associated with increased levetiracetam concentrations include general anemia, neutropenia, and significant drowsiness (somnolence). These toxicities can sometimes occur even when the drug concentration is within the therapeutic interval.

Gabapentin is another of the newer antiepileptic drugs introduced in the early 1990s for treatment of seizures. Since its introduction, it has also found use for treatment of neuropathic pain, restless leg syndrome, anxiety, bipolar disorder, and insomnia. In epilepsy, it is often used for the treatment of drug-resistant partial seizures. Gabapentin does not undergo any hepatic metabolism and is cleared almost completely by renal excretion—the elimination half-life is approximately 5−7 hours in patients with normal renal function. The generally reported therapeutic interval for C_0 concentrations is 2−20 μg/mL [8]. Toxic effects of high gabapentin

concentration include extreme fatigue, drowsiness, dizziness, and ataxia (loss of full control of body movements).

Oxcarbazepine is a prodrug that is metabolized to a biologically active metabolite, 10-hydroxy-10,11-dihydrocarbamazepine, known more commonly as monohydroxycarbamazepine (MHD). MHD is responsible for the antiseizure activity of the drug, and it is the component that is measured in the blood for TDM. The half-life for MHD is longer than that of oxcarbazepine (8−10 hours vs. 1−2.5 hours). The therapeutic interval for MHD is reported generally as 3−35 μg/mL [9]. Toxicity associated with MHD includes dizziness, drowsiness, GI toxicity, tremor, ataxia, and abnormal gait. These toxicities can sometimes occur even when the drug concentration is within the therapeutic interval.

Topiramate is a newer broad-spectrum antiepileptic drug that has also been used for treatment of seizures including those from Lennox−Gastaut syndrome in children and also bipolar disorder, cluster headaches, and trigeminal neuralgia. It exerts antiseizure activity through multiple mechanisms and is notable in that it blocks seizure activity without also increasing the potential for inducing seizures with increasing concentration. The serum half-life for topiramate is 20−30 hours, and the therapeutic interval is generally listed as 5−20 μg/mL. However, there is significant overlap in blood concentrations between the responders and the nonresponders. Toxicity associated with topiramate includes poor cognition and physical side effects such as fatigue, somnolence, joint pain, and GI toxicity.

Transplantation

Transplantation is an extremely complex medical procedure, and immunosuppression is central to controlling the body's response to the transplanted organ to ensure a successful xenograft. Management of these powerful drugs requires the balance of ensuring enough drug exposure to prevent rejection while keeping the concentration of the drug low enough to avoid toxic effects. The use of immunosuppression has evolved significantly over the years from the mid-20th century until now. With current protocols, TDM is an essential tool in the process, allowing rapid titration of blood concentrations of the drug to maximize immunosuppression and avoid acute rejection while minimizing adverse events from exposure to the drug.

Tacrolimus is a calcineurin inhibitor and the most widely used immunosuppressive drug used in transplantation. It can be administered either intravenously or orally, and exhibits significant interindividual variability. This drug exerts its immunosuppressive effects by blocking the synthesis of cytokines and inflammatory mediators. Tacrolimus is monitored in whole blood rather than serum or plasma, and a general therapeutic interval is reported as 5−15 ng/mL for C_0 concentrations [10], although, in practice, target concentration ranges are more narrow and dependent on the type of organ transplanted as well as the time from transplantation. Adverse effects of elevated concentrations of tacrolimus in blood include nephrotoxicity, neurotoxicity, hypertension, and nausea.

Cyclosporine A (CsA) is a calcineurin inhibitor that has been available for longer than tacrolimus, but is not as widely used. It exerts its immunosuppressive effects by blocking the activation of T lymphocytes. CsA is available both in intravenous and oral forms, with variable absorption and distribution; it is also highly protein bound. CsA is found in both plasma and red blood cells, but measurement of the drug primarily occurs in whole blood samples, as the drug distributes into red cells in vitro after collection as the temperature falls. The general target interval for C_0 concentrations is 100−400 ng/mL [11], although the target in clinical practice is dependent on multiple factors including type of transplant, time from transplantation, method of analysis, and coadministered drugs. There is some evidence that C_2 (peak concentration 2 hours after administration) is more closely correlated with clinical outcomes [12]. CsA is metabolized in the liver by CYP3A4, and the drug demonstrates significant interindividual pharmacokinetic variability and is susceptible to drug−drug interactions. Adverse effects of high CsA concentrations include renal toxicity, as well as liver or CNS effects.

Sirolimus is an IL-2 inhibiting drug that targets the mammalian target of rapamycin (mTOR) receptor, which interrupts the cell cycle. Sirolimus is primarily metabolized by CYP3A4/5, so it exhibits significant pharmacokinetic variability and is susceptible to drug−drug interactions. The half-life of sirolimus is relatively long (48−72 hours) compared with other immunosuppressants, so it takes longer to achieve steady-state concentrations after initiation of therapy or with a dosage change. There is also a prodrug available called temisirolimus, which is metabolized to sirolimus via CYP3A4. Sirolimus inhibits T-cell activation and proliferation by inhibiting mTOR, which leads to cell cycle interruption. Generally, the therapeutic interval for sirolimus is 4−12 ng/mL, but as with the other drugs, target concentrations are dependent on transplant type and comedications, as well as other physiological parameters [13]. Sirolimus is measured in whole blood samples rather than serum or plasma due to its significant distribution into red blood cells. The benefit of this drug is that it does not have renal toxicity, making it appealing for use in kidney transplantation. However, significant toxic effects are still possible including leukopenia, thrombocytopenia, and hypercholesterolemia.

Everolimus is a structural analog of sirolimus, and as such, they share the same target and mechanism of action (mTOR inhibition). The primary difference between

everolimus and sirolimus is that everolimus has a shorter half-life (18−36 hours) and its toxic effects are less significant. Everolimus is metabolized via CYP3A4, and is subject to drug−drug interactions, including inhibition by coadministration of cyclosporine. Therefore it is still important to manage this drug using TDM for optimal therapeutic effect while minimizing toxicities. Everolimus is measured in whole blood samples for TDM. The target therapeutic interval is 3−15 ng/mL, with modifications based on coadministered medications, transplant type, and time from transplantation [14].

Mycophenolic acid (MPA) is an inhibitor of inosine monophosphate dehydrogenase (IMPDH), which inhibits cell group by inhibition of purine synthesis. Purine synthesis catalyzed by IMPDH is the rate-limiting step in proliferating lymphocytes, so inhibition by MPA significantly decreases the response of the activated T cells. MPA is the active compound; however, the drug is actually administered as a prodrug such as mycophenolate mofetil (CellCept) or mycophenolate sodium (Myfortic), both of which are extended release formulations. The drug is used in conjunction with either calcineurin or mTOR inhibitors as adjuvant therapy for immunosuppression. MPA has a relatively short half-life (9−18 hours) and is not distributed into red blood cells like the other commonly used immunosuppressant drugs. Important differences for MPA include (1) the measurement of the drug is in plasma rather than whole blood; and (2) a peak or C_0 measurement is not sufficiently correlated with area under the curve (AUC), so a single measurement is not sufficient. It is recommended that AUC with limited sampling be used for TDM of this compound [15]. The standard AUC target for transplantation is 30−60 mg/h.L. When a single measurement is used, a predose range of 1−4 ug/mL is considered therapeutic, with modification depending on the coadministered drug. The primary toxicity for MPA is gastrointestinal toxicity—there is some debate of whether TDM is really needed for MPA or whether the dose can just be reduced when GI toxicity is observed.

Cardioactive drugs

Historically, in cardiology, certain therapeutic agents have been used to control arrhythmia in patients with abnormal heart physiology. However, these historic therapeutic agents displayed two important tendencies. First, each of the agents demonstrated significant pharmacokinetic variability. Second, while these drugs can control arrhythmia when the dose is optimized, they can also cause arrhythmia when concentrations in blood are too high. Based on these characteristics, it was recognized that TDM is a pivotal tool in optimal management of these agents, as toxic effects can even appear at concentrations just slightly higher than the target range.

Procainamide is an antiarrhythmic drug used for control of both atrial and ventricular arrhythmias, and *N-acetylprocainamide* (NAPA) is the primary metabolite that also elicits antiarrhythmic activity (via a different mechanism). The primary mechanism of action for procainamide is through Na^+ channel blockade. The half-life of procainamide is approximately 3−4 hours and the half-life for NAPA is closer to 6 hours, and both compounds exhibit significant interindividual variability. The therapeutic target for procainamide is 4−8 μg/mL, and for NAPA, it is 10−20 μg/mL; when considered together, the therapeutic target interval for procainamide and NAPA combined is 5−30 μg/mL [16]. Adverse effects associated with high concentrations of procainamide and NAPA include hypotension, induced arrhythmia, and widening of QRS intervals (associated with chronically high exposure).

Lidocaine is a drug commonly used for emergent treatment of acute or life-threatening arrhythmias. Lidocaine is also a Na^+ channel blocker, but it is more weakly bound than the procainamide. It is highly protein-bound (to α-1-acid glycoprotein) and is metabolized in the liver. Since lidocaine is used in acutely ill patients and α-1-acid glycoprotein is an acute phase reactant, it is common for the free/unbound fraction of the drug to change rapidly along with the patient condition, thus changing the effective concentration of the drug. However, despite this, the free fraction of lidocaine is not commonly monitored. In addition, the hepatic metabolism of the drug means that the pharmacokinetics can be significantly impacted by changes in liver perfusion or liver injury, which is particularly pertinent in an acutely ill population. The therapeutic interval for lidocaine is reported as 1.5−5 μg/mL [17]. Toxic effects from high lidocaine concentrations include bradycardia, hypotension, and CNS dysfunction.

Quinidine is used for both atrial and ventricular arrhythmias, but it is less commonly administered due to the GI side effects present even when blood concentrations are within the therapeutic range. It is a stronger Na^+ channel blocker than lidocaine. Blood concentrations are best measured as C_0 levels, and the therapeutic interval is reported as 2−5 μg/mL [18]. Concentrations greater than 5 μg/mL are associated with hypotension, ventricular tachycardia, or fibrillation, and cinchonism and QT interval elongation on an electrocardiogram. Interestingly, the QT interval elongation is a desired effect for patients with Brugada Syndrome, making it a preferred treatment in that population [19]. Several metabolites of quinidine have been shown to have biologic activity, but they are not routinely monitored.

Digoxin is a cardiac glycoside used in treatment of cardiac arrhythmia and also congestive heart failure.

Digoxin demonstrates multiple mechanisms of action, including regulation of cation flux in myocardial cells as well as sympathetic nervous and endocrine systems that impact cardiac function. One of the primary toxic effects of the drug at higher than therapeutic concentrations is cardiac arrhythmia that is difficult to differentiate from the clinical indication for the drug in the first place. The therapeutic interval for digoxin is reported as 0.8−2.0 ng/mL [20]. Toxic symptoms in addition to arrhythmia include GI toxicity and neurologic symptoms. Digoxin assays can encounter significant interference from digoxin-like immunoreactive factors or Digibind (Fab fragments specific for digoxin used for treatment of overdose).

Psychoactive drugs

In the treatment of psychiatric disorders such as depression or schizophrenia, pharmacotherapy is commonly managed by titration to clinical effect. This approach is taken because of the lack of robust relationship between the dose administered (or blood concentration) and the clinical response to therapy—the evaluation of response to therapy is often subjective in nature. It can vary based on the variability of how the patients communicate their experience and also the way the clinicians perceive their interaction with the patient. However, for some drugs, the adverse effects of the drug can be objectively measured and correlated with blood concentration of the drug. In these cases, TDM is a useful tool for psychoactive drug management.

Tricyclic antidepressants include amitriptyline, nortriptyline, imipramine, desipramine, and clomipramine, and are used for treatment of major depression. While a robust relationship between blood concentration and therapeutic clinical effect is difficult to define, there is a well-described relationship between the blood concentrations of the tricyclic antidepressant drugs and the toxic effects (life-threatening cardiac arrhythmia). It is because of this significant toxicity and potential for suicide using the prescribed therapeutic drug that tricyclic antidepressants are not widely used at present. However, in certain patient populations, these drugs are still preferred, and when used, TDM is necessary to ensure patient safety. Blood concentrations greater than 500 ng/mL are generally reported as toxic and are associated with increased risk of cardiac arrhythmia [21]. For other antidepressant drugs such as selective serotonin reuptake inhibitors and selective norepinephrine reuptake inhibitors, the toxic side effects of the drugs are not so severe, and without a well-defined therapeutic interval, TDM of these drugs is not commonly performed.

Lithium is a drug that is widely used for treatment of bipolar disorder; specifically, it is used for control of the manic phase in the disorder. The blood concentration of lithium can be affected by changes in other physiologically relevant electrolytes. For instance, increased intake of sodium can enhance lithium excretion, while decreased physiologic sodium concentration can lead to reduced lithium excretion. Lithium is cleared from circulation only by the kidneys, so decreased renal function can lead to toxic accumulation of the drug. The therapeutic interval for lithium is generally reported as 0.4−1.2 mmol/L, with toxic effects seen at concentrations greater than 1.5 mmol/L [22]. Clinically significant side effects of high lithium concentrations include renal failure and excessive water and electrolyte loss. Chronic lithium administration can also affect thyroid function (up to 35% of patients treated with lithium develop hypothyroidism), so thyroid function should be monitored regularly for patients on lithium therapy [23].

More recently, TDM of antipsychotic drugs has emerged as an area of interest for TDM in psychiatry—particularly for monitoring adherence to therapy. While the same challenges exist for these drugs as in antidepressants in terms of drug exposure predicting clinical success in treatment, there is growing interest in using drug monitoring to assess adherence to therapeutic regimens. Common drugs subjected to TDM for this purpose include clozapine, olanzapine, risperidone, quetiapine, and aripiprazole. Of course, to assess whether a patient is taking the drug does not require a blood sample or TDM—one recent study has demonstrated the potential of using urine metabolites for assessment of adherence in patients treated with aripiprazole [24]. However, TDM is still not a standard of practice for most psychiatric patients.

Infectious disease

In the treatment of infectious diseases, the biologic activity of the drug is directed against the microorganism responsible for the infection and not toward the person taking the drug. Given this mechanism of action, the therapeutic index is quite wide and the while many of the drugs have significant pharmacokinetic variability and it is important to have the drug concentration (exposure) higher than the minimal effective concentration against the microorganism, the lack of toxic effects against the host/patient allows for a high enough dose to ensure sufficient blood concentrations for efficacious treatment for patients regardless of pharmacokinetic variability. However, some antimicrobial drugs do have both significant pharmacokinetic variability and significant clinical toxicity for the patient. In these cases, TDM is a valuable tool for optimal management of patients on antimicrobial therapy.

Aminoglycoside antibiotics include gentamicin and tobramycin (effective against both Gram-positive and Gram-negative bacteria), as well as amikacin (effective primarily against Gram-negative bacteria). Aminoglycosides

are characterized by their short half-lives (1−2 hours) and lack of oral bioavailability—they are all administered intravenously. The aminoglycoside antibiotics are renally cleared and exhibit significant interindividual pharmacokinetic variability. Traditional dosing of these drugs includes dosing every 6−8 hours, with peak monitoring target concentrations of 5−10 μg/mL for gentamicin and tobramycin and 20−25 μg/mL for amikacin [25]. However, due to the antibiotic residual effect [26], where the antibiotic effects of the drug are related to the peak exposure concentration versus toxic effects that are related to the duration of exposure, daily dosing (or pulse dosing) is more commonly applied in current regimens. For this therapeutic paradigm, all of the doses for a day are combined into a single administered dose for the entire day. When this approach is taken, the peak therapeutic target intervals established in the literature are no longer applicable, and TDM is performed using C_0 monitoring primarily to ensure that the drug is being cleared. In this approach, C_0 levels should be less than 2 μg/mL. Significant toxic effects are commonly observed when these drugs are used, including irreversible ototoxicity and renal toxicity in the form of renal tubular damage.

Vancomycin is a glycopeptide antibiotic used for treatment of many Gram-positive bacterial infections and also demonstrates treatment efficacy for some Gram-negative bacteria as well. It works particularly well for methicillin-resistant staph aureus infection. Similar to the aminoglycosides, vancomycin does not have oral bioavailability and it is administered as an intravenous infusion. The half-life of vancomycin is 4−6 hours, and it is typically given twice per day. It is cleared via the kidneys with significant interindividual variability—TDM is performed using trough levels with a target concentration interval of 10−20 μg/mL [27]. Toxic effects of vancomycin include ototoxicity and nephrotoxicity, although these toxic effects are more commonly seen when the drug is coadministered with drugs that share a similar toxicity profile such as the aminoglycoside antibiotics.

Voriconazole is a triazole antifungal drug that is used to treat invasive fungal infections (such as candidiasis or aspergillosis) and also for prophylaxis in severely immunocompromised patients, such as advanced HIV patients or those with leukemia or in preparation for bone marrow transplantation. Voriconazole is administered both orally and intravenously, with an oral bioavailability of 96%. It is metabolized in the liver, and demonstrates significant interindividual pharmacokinetic variability and also is susceptible to a number of drug−drug interactions. The reported therapeutic target interval for voriconazole is 1.0−5.5 μg/mL, and adverse side effects include GI toxicity, headache, peripheral edema, and visual disturbances. The primary toxicity associated with blood concentrations greater than 6.0 μg/mL is liver toxicity [28].

While the antimicrobial drugs previously discussed are those most commonly managed using TDM, there are some emerging applications in the area of infectious disease. As the drug posaconazole is more frequently used to treat antifungal infections, there are those that suggest it should be monitored similarly to voriconazole and for the same reasons [29]. In addition, there has been an increase recently regarding the application of TDM for antituberculosis drugs. Beta-lactam drugs exhibit significant pharmacokinetic variability, and there are a number of studies that suggest TDM may be of benefit in patients administered those drugs as well.

Oncology

In the treatment of cancer, most drugs are dosed based on body surface area normalization (e.g., mg/m^2), and then, the patients are monitored for significant toxicity or tumor response—this is similar to the "titrate to clinical effect" scheme described earlier. However, in cancer treatment, there is an additional consideration of maximum tolerated dose (MTD). Since the drugs are known to be toxic, part of the clinical trials during drug development involves assessment of dose limiting toxicity relative to the administered dose in a small number of patients, in order to determine the maximum amount of drug that should be given. When patients are receiving the MTD and not exhibiting symptoms of toxicity, treatment is continued without dose adjustment until it is determined whether the treatment is affecting the cancer. However, it is important to note that many chemotherapy drugs have significant pharmacokinetic variability, so that when MTD is given to the patient and no toxic effects are observed, it is not certain that the patient is getting just the right dose, it may be that the patient is actually getting less drug than needed (or that they can tolerate). Based on this, some have advocated for the concept of maximum tolerated exposure [30], which would require blood concentration measurements. TDM is not routine in the management of chemotherapy for the most part.

Methotrexate is a folic acid antagonist drug that initially was used as an immunosuppressant drug in transplantation, but has found more widespread use in the treatment of malignancy or autoimmune disease. It exerts its activity by inhibiting the enzyme dihydrofolate reductase, which decreases concentrations of tetrahydrofolate and ultimately leads to the cell death. The adverse effects can be reversed with leucovorin (a folic acid derivative), which restores tetrahydrofolate-dependent synthesis of pyrimidines and leads to cell rescue. As with many other drugs that require TDM, methotrexate exhibits significant interindividual variability and is susceptible to drug−drug interactions. In the treatment of autoimmune disorders, low-dose methotrexate is used, and typically, TDM is not

needed for management of these patients. However, in treatment of malignancy, the risk of adverse effects on normal cells is much greater and the drug is commonly used with leucovorin to prevent damage to noncancer cells. The amount of leucovorin needed is dependent on the rate of methotrexate clearance, so methotrexate concentrations are measured over multiple days. The target concentration intervals are 5−10 μM at 24 hours postdose, 0.5−1.0 μM at 48 hours postdose, and 0.05−0.1 μM at 72 hours postdose [31]. When methotrexate concentrations are greater than those time-dependent target intervals, addition leucovorin is needed. The adverse effects of high methotrexate levels include hepatotoxicity, leukopenia, and ulcerative stomatitis.

Busulfan is an alkyl sulfonate drug that functions as a nonspecific alkylating neoplastic agent. It inhibits the growth of cancerous cells by alkylating DNA. Busulfan was initially approved as a treatment for chronic myeloid leukemia (CML); however, it is currently primarily used as a conditioning agent along with cyclophosphamide, fludarabine, or total body irradiation prior to bone marrow transplantation. Busulfan can be given as either an oral or intravenous formulation; the kinetics are more predictable when given intravenously. TDM-based dosing of busulfan is not based on a single measurement, but is instead based on 6-hour AUC. The general target AUC is between 900 and 1500 μmol.min. When the AUC is less than 900, the risk of incomplete bone marrow ablation is increased, along with the chances of graft rejection or graft-versus-host disease—in these cases, the dose should be increased. AUCs greater than 1500 are associated with liver toxicity known as sinusoidal obstructive syndrome, also known as venoocclusive disease, which can result in significant organ damage and poor clinical outcomes [32]. In some newer treatment protocols, once-daily dosing of busulfan is used rather than four times per day dosing, with higher AUC targets [33].

While TDM has been sparsely utilized in oncology with the exception of methotrexate and busulfan, there is an emerging body of evidence suggesting that it can be a useful tool for managing other chemotherapy drugs. There is a growing body of scientific literature suggesting that phenotpying for the thiopurine drug mercaptopurine by measuring the drug metabolites (TDM) is useful for managing the drug; however, it is just as common or more so to obtain the genotype prior to dosing or determine the enzymatic activity directly through the bioassay of the red blood cells for a patient [34,35]. There are a large number of studies supporting TDM in the scientific literature for optimization of 5-fluorouracil (5-FU) management, including one of the few randomized control trials published in 2008 [36]. However, despite the large number of studies suggesting the utility of 5-FU monitoring, it has not become a standard of practice. More recently, clinical studies have demonstrated the potential for TDM to reduce toxicity in patients receiving taxanes (docetaxel and paclitaxel) for treatment of cancer [37]. Finally, several studies have suggested that TDM for management of imatinib in the settings of CML [38] and gastrointestinal stromal tumors [39] has potential to affect patient outcomes positively. However, despite the emerging amount of evidence, TDM remains far from routine implementation in oncology settings.

References

[1] P.N. Patsalos, D.J. Berry, B.F. Bourgeois, J.C. Cloyd, T.A. Glauser, S.I. Johannessen, et al., Antiepileptic drugs-best practice guidelines for therapeutic drug monitoring: a position paper by the sub-commission on therapeutic drug monitoring, ILAE Commission on Therapeutic Strategies, Epilepsia 49 (2008) 1239−1276.

[2] M.J. Brodie, A. Richens, A.W. Yeun, Double-blind comparison of lamotrigine and carbamazepine in newly diagnosed epilepsy. UK Lamotrigine/Carbamaepin Monotherapy Trial Group, Lancet 345 (1995) 476−479.

[3] H.E. Booker, K. Hosokowa, R.R. Burdette, B. Darcey, A clinical study of serum primidone levels, Epilepsia 11 (1970) 395−402.

[4] D. Schmidt, I. Einicke, F.T. Haenel, The influence of seizure type on the efficacy of plasma concentrations of phenytoin, phenobarbital, and carbamazepine, Arch. Neurol. 43 (1986) 263−265.

[5] O. Henriksen, S.I. Johannessen, Clinical and pharmacokinetic observations on sodium valproate − a 5-year follow-up study in 100 children with epilepsy, Acta Neurol. Scand. 65 (1982) 504−523.

[6] R.G. Morris, A.B. Black, A.L. Harris, A.B. Batty, B.C. Sallustio, Lamotrigine and therapeutic drug monitoring: retrospective survey following the introduction of a routine service, Br. J. Clin. Pharmacol. 46 (1998) 547−551.

[7] I.E. Leppik, J.O. Rarick, T.S. Walczak, T.A. Tran, J.R. White, R. J. Gumnit, Effective levetiracetam doses and serum concentrations: age effects, Epilepsia 43 (Suppl 7) (2002) 240.

[8] J. Sivenius, R. Kalviainen, A. Ylinen, P. Riekkinen, A double-blind study of gabapentin in the treatment of partial seizures, Epilepsia 32 (1991) 539−542.

[9] S. Striano, P. Striano, P. Di Nocera, D. Italiano, C. Fasiella, P. Ruosi, et al., Relationship between serum mono-hydroxy-carbazepine concentrations and adverse effects in patients with epilepsy on high-dose oxcarbazepione therapy, Epilepsy Res. 69 (2006) 170−176.

[10] Y.P. Agarwal, M. Cid, S. Westgard, T.S. Parker, R. Jaikaran, D. M. Levine, Transplant patient classification and tacrolimus assays: more evidence of the need for assay standardization, Ther. Drug. Monit. 36 (2014) 706−709.

[11] M. Oellerich, V.W. Armstrong, B. Kahan, L. Shaw, D.W. Holt, R. Yatscoff, et al., Lake Louise Consensus Conference on cyclosporine monitoring in organ transplantation: report of the consensus panel, Ther. Drug. Monit. 17 (1995) 642−654.

[12] L. Pape, A. Lehnhardt, K. Latta, J.H. Ehrich, G. Offner, Cyclosporin A monitoring by 2-h levels: preliminary target levels

in stable pediatric kidney transplant recipients, Clin. Transplant. 17 (2003) 546–548.
[13] R.G. Morris, P. Salm, P.J. Taylor, F.A. Wicks, A. Theodossi, Comparison of the reintroduced MEIA assay with HPLC-MS/MS for the determination of whole-blood sirolimus from transplant recipients, Ther. Drug. Monit. 28 (2006) 164–168.
[14] V.H. Mabasa, M.H. Ensom, The role of therapeutic monitoring of everolimus in solid organ transplantation, Ther. Drug. Monit. 27 (2005) 666–676.
[15] L.M. Shaw, I. Nowak, Mycophenolic acid: measurement and relationship to pharmacologic effects, Ther. Drug. Monit. 17 (1995) 685–689.
[16] M. Wierzchowiecki, K. Jasinski, R. Trojanowicz, R. Ochotny, T. Tomaszkiewicz, Pharmacokinetic and pharmacodynamics studies of procainamide given intermittently intravenously in patients with sever ventricular arrhythmias, Cor Vasa 20 (1978) 176–183.
[17] E.L. Alderman, R.E. Kerber, D.C. Harrison, Evaluation of lidocaine resistance in man using intermittent large-dose infusion techniques, Am. J. Cardiol. 34 (1974) 342–349.
[18] H. Halkin, Z. Vered, P. Milman, B. Rabinowitz, H.N. Neufeld, Steady-state serum quinidine concentration: role in prophylactic therapy following acute myocardial infarction, Isr. J. Med. Sci. 15 (1979) 583–587.
[19] P.L. Pellegrino, M. Di Biase, N.D. Brunetti, Quinidine for the management of electrical storm in an old patient with Brugada syndrome and syncope, Acta Cardiol. 68 (2013) 201–203.
[20] J.R. Horn, D.B. Christensen, P.A. deBlaquiere, Evaluation of a digoxin pharmacokinetic monitoring service in a community hospital, Drug. Intell. Clin. Pharm. 19 (1985) 45–52.
[21] S.H. Preskorn, R.C. Dorey, G.S. Jerkovich, Therapeutic drug monitoring of tricyclic antidepressants, Clin. Chem. 34 (1988) 822–828.
[22] C.S. Frings, Lithium monitoring, Clin. Lab. Med. 7 (1987) 545–550.
[23] B. Shine, R.F. McKnight, L. Leaver, J.R. Geddes, Long-term effects of lithium on renal, thyroid, and parathyroid function: a retrospective analysis of laboratory data, Lancet 386 (2015) 461–468.
[24] J. McEvoy, R.A. Millet, K. Dretchen, A.A. Morris, M.J. Corwin, P. Buckely, Quantitative levels of aripiprazole parent drug and metabolites in urine, Psychopharmacology (Berl.) 231 (2014) 4421–4428.
[25] E.J. Begg, M.L. Barclay, C.M. Kirkpatrick, The therapeutic drug monitoring of antimicrobial agents, Br. J. Clin. Pharmacol. 52 (Suppl 1) (2001) 35S–43S.
[26] M.S. Stankowicz, J. Ibrahim, D.L. Brown, Once-daily aminoglycoside dosing: an update on current literature, Am. J. Health Syst. Pharm. 72 (2015) 1357–1364.
[27] R. Kullar, S.N. Leonard, S.L. Davis, G. Delgado Jr, J.M. Pogue, K.A. Wahby, et al., Validation of the effectiveness of a vancomycin nomogra in achieving target trough concentrations of 15–20 mg/L suggested by the vancomycin consensus guidelines, Pharmacotherapy 31 (2001) 441–448.
[28] H.Y. Chu, R. Jain, H. Xie, P. Pottinger, D.N. Fredricks, Voriconazole therapeutic drug monitoring: retrospective cohort study of the relationship to clinical outcomes and adverse events, BMC Infect. Dis. 13 (2013) 105.
[29] M.R. Ananda-Rajah, A. Grigg, M.A. Slavin, Making sense of posaconazole therapeutic drug monitoring: a practical approach, Curr. Opin. Infect. Dis. 25 (2012) 605–611.
[30] J.H. Beumer, E. Chu, S.J. Salamone, Body-surface area-based chemotherapy dosing: appropriate in the 21st century? J. Clin. Oncol. 30 (2012) 3896–3897.
[31] E. Dombrowsky, B. Jayaraman, M. Narayan, J.S. Barrett, Evaluating performance of a decision support system to improve methotrexate pharmacotherapy in children and young adults with cancer, Ther. Drug. Monit. 33 (2011) 99–107.
[32] R. Malar, F. sjoor, K. Rentsch, M. Hassan, T. Gungor, Therapeutic drug monitoring is essential for intravenous busulfan therapy in pediatric hematopoietic stem cell recipients, Pediatr. Transpl. 15 (2011) 580–588.
[33] S.G. Ryu, J.H. Lee, S.J. Choi, J.H. Lee, Y.S. Lee, M. Seol, et al., Randomized comparison of four-times-daily versus once-daily intravenous busulfan in conditioning therapy for hematopoietic cell transplantation, Biol. Blood Marrow Transpl. 13 (2007) 1095–1105.
[34] R. Weinshilboum, Inheritance and drug response, N. Engl. J. Med. 348 (2003) 529–537.
[35] R.M. Weinshilboum, S.L. Sladek, Mercaptopurine pharmacogenetics: monogenic inheritance of erythrocyte thipurine methyltransferase activity, Am. J. Hum. Genet. 32 (1980) 651–652.
[36] E. Gamelin, R. Delva, J. Jacob, Y. Merrouche, J.L. Raoul, D. Pezet, et al., Individual fluorouracil dose adjustment based on pharmacokinetic follow-up compared with conventional dosage: results of a multicenter randomized trial of patients with metastatic colorectal cancer, J. Clin. Oncol. 26 (2008) 2099–2105.
[37] S. Kraff, A.J. Nieuweboer, R.H. Mathijssen, F. Baty, A.J. de Graan, R.H. van Schaik, et al., Pharmacokinetically based dosing of weekly paclitaxel to reduce drug-related neurotoxicity based on a single sample strategy, Cancer Chemother. Pharmacol. 75 (2015) 975–983.
[38] H. Yu, N. Steeghs, C.M. Mijenhuis, J.H. Schellens, J.H. Beijnen, A.D. Huitema, Practical guidelines for therapeutic drug monitoring of anticancer tyrosine kinase inhibitors: focus on the pharmacokinetic targets, Clin. Pharmacokinet. 53 (2014) 305–325.
[39] A. Sawaki, K. Inaba, S. Nomura, H. Kanie, T. Yamada, K. Hayashi, et al., Imatinib plasma levels during successful long-term treatment of metastatic-gastrointestinal stromal tumors, Hepatogastroenterology 61 (2014) 1984–1989.

Self-assessment questions

1. Criteria for effective TDM include:
   - **a.** list of concomitant medications for the patient
   - **b.** the relationship between the drugs dose and the blood concentration must be highly variable and/or not predictable
   - **c.** the relationship between blood concentration and clinical effect must be well defined
   - **d.** a and b
   - **e.** b and c
2. In addition to adjustment of drug dosage for optimizing therapy, TDM can also be useful for:
   - **a.** ensuring a complete medical record
   - **b.** assessment of medication adherence
   - **c.** risk management
   - **d.** nothing; it is only effective for dose optimization
3. Timing of sample collection is most important for drugs with which of the following?
   - **a.** Very short half-lives
   - **b.** Variable bioavailability
   - **c.** Active metabolites
   - **d.** No recognized circadian rhythm effects
   - **e.** Very long half-lives
4. Which one of the following classes of drugs is commonly monitored at the peak and the trough?
   - **a.** Antiepileptics
   - **b.** Aminoglycosides
   - **c.** Cardioactive drugs
   - **d.** Immunosuppressants
   - **e.** Tricyclic antidepressants
5. Which of the following immunosuppressive drugs requires multiple time points for effective TDM?
   - **a.** Cyclosporine A
   - **b.** Tacrolimus
   - **c.** MPA
   - **d.** Sirolimus
   - **e.** Everolimus
6. Which of the following drugs is commonly monitored in oncology?
   - **a.** 5-FU
   - **b.** Docetaxel
   - **c.** Doxorubicin
   - **d.** Methotrexate

Answers

1. e
2. b
3. a
4. b
5. c
6. d

Chapter 52

Toxicology and the clinical laboratory

Loralie J. Langman[1] **and Paul J. Jannetto**[2]

[1]*Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, United States,* [2]*Clinical and Forensic Toxicology Laboratory, Mayo Clinic, Rochester, MN, United States*

Learning objectives

After reviewing this chapter, the reader will be able to:

- Define the different toxidromes and describe some of the key signs and symptoms and agents that can cause each one.
- For each of the following cellular hypoxia causative agents, describe the agent, the toxic effects, the treatment, and the method(s) of analysis, including specimen requirements: carbon monoxide, cyanide, and methemoglobin-forming agents.
- For each of the following alcohols, describe the pharmacological action, the metabolites, the antidote, other biomarkers of exposure, and the method(s) of analysis, including specimen requirements: ethanol, isopropanol, methanol, and ethylene glycol.
- For each of the following analgesics, describe the toxic effects of overdose, the dose and time of symptom onset, the metabolizing enzyme and metabolite, the antidote, and the method(s) of analysis, including specimen requirements: acetaminophen and salicylate.
- For the following drugs, describe the pharmacological response, significant metabolic pathways, clinical uses if any, the antidote, and the screening and confirmatory tests: amphetamines/methamphetamines, barbiturates, benzodiazepines, cannabinoid, cocaine (COC), marijuana, opioids, phencyclidine, synthetic opioids, and tricyclic antidepressants.
- For the following metals be able to recognize common physical signs/symptoms of toxicity, identify appropriate laboratory tests, specimen types, the mechanism of toxicity, and the treatment or management of toxicity: lead, arsenic, mercury, cadmium, and chromium.

Background

Toxicology is an extensive, multidisciplinary science with an overarching goal to determine the effects of chemicals on living systems. Numerous potential toxins can cause harm, including over-the-counter or prescription drugs, herbal medications, common household products, environmental agents, occupational substances, drugs of abuse, and chemical terrorism threats. Unfortunately, there is no specific antidote or treatment for most toxins; therefore supportive care is the most appropriate intervention [1]. In clinical practice, it is also not required or realistic to test for all of the potential toxins that a person may have been exposed to. In fact, approximately 25 substances account for 80% or more of the cases treated in emergency departments [2]. Therefore the scope of clinical toxicology testing depends on the local pattern of drug abuse, the resources of the laboratory, and the clinical need of the institution to provide care for its patients.

In the end, all compounds have the potential to be toxic, and the amount a person is exposed to or consumes determines if the compound is poisonous or not. Poisonings most commonly involve products or chemicals readily available to the victim. One way to evaluate the toxicity of a substance is to look at the median lethal dose or lethal dose 50% (LD_{50}). The LD_{50} is the amount of a substance that is required to cause death in half of the exposed population; the LD_{50} is typically determined in animal studies. In the end, the basic pharmacokinetic principles discussed previously in Chapter 49, The porphyrias: fundamentals and laboratory assessment, affect the extent of the toxicity. In addition, preexisting organ dysfunction, genetic factors, health status, or preexposure to compounds that can alter the biotransformation of the compound can all influence the toxic effects. Furthermore, the duration and frequency of the exposure can also affect the onset and severity of toxicity.

Laboratory methods

A variety of different analytical approaches are commonly utilized in toxicology laboratories including rapid spot tests, immunoassays, chromatographic methods [e.g., high-performance liquid chromatography (HPLC) or gas chromatography (GC)], chromatographic methods combined with mass spectrometry (MS; e.g., GC−MS, GC−MS/MS, LC−MS, or LC−MS/MS), and high-resolution mass spectrometry (e.g., HRMS, TOF, and Orbitrap) [3]. One key

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00052-1

issue in selecting an appropriate methodology to use in the clinical laboratory is the speed of analysis, or turnaround time (TAT) required. A drug test that requires several hours to complete the analysis or a test that is not available 24 hours a day/7 days a week is less useful to physicians treating a poisoned patient in the Emergency Department.

As a result, when a patient presents with a potential toxicology-related health problem, the diagnosis is often determined by (1) reviewing the case history; (2) physical exam findings; (3) ancillary tests (e.g., electrocardiogram); and (4) applying a rational and evidence-based approach to laboratory testing. Screening procedures are designed for a relatively rapid and generally qualitative detection of a particular drug or drug class. In general, screening tests have adequate clinical sensitivity but may not be highly specific. Therefore a negative result from a screening procedure may rule out, with reasonable certainty, the presence of a clinically significant concentration of a particular analyte. However, a positive result should be considered only "presumptive positive," and it should be confirmed by an alternate procedure with greater specificity.

Toxidromes

Toxic syndromes (toxidromes) are a group of signs and symptoms that occur together and characterize a particular abnormality or class of poisons. The most common toxidromes include (1) anticholinergic; (2) cholinergic; (3) opioid; (4) sedative-hypnotic; and (5) sympathomimetic (Table 52.1). However, many poisonings involve more than one agent, and some toxidromes have overlapping features. For example, anticholinergic findings are very similar to sympathomimetic findings, with an exception being the effects on sweat glands: anticholinergic agents produce warm, flushed dry skin, but sympathomimetic agents produce diaphoresis. Toxidrome findings may also be affected by individual variability, comorbidity, and coingestants. When accurately identified, the toxidrome may provide valuable information for a quicker diagnosis and subsequent treatment, although the many limitations impeding acute toxidrome diagnosis must be carefully considered.

Cholinergic toxidrome

Acetylcholine is a neurotransmitter found throughout the central nervous system (CNS), including (1) the sympathetic and parasympathetic autonomic ganglia; (2) the postganglionic parasympathetic nervous system; and (3) the skeletal muscle motor end plate. The symptoms associated with cholinergic poisoning are due to the inhibition of acetylcholinesterase (AChE), which results in the buildup of excessive levels of acetylcholine at the synapses or neuromuscular junctions and varies depending on the balance of nicotinic and muscarinic stimulation. Activating muscarinic receptors stimulates or inhibits cellular function at visceral smooth muscle, cardiac muscle, and secretory glands. Two mnemonics have been developed to help remember cholinergic clinical effects: DUMB BELLS (diarrhea, urination, miosis, bradycardia, and bronchorrhea-bronchoconstriction, emesis, lacrimation, lethargy, and sweating-salivation) [4] and SLUDGE [salivation, lacrimation, urination, defecation, gastrointestinal (GI) distress, and emesis-eye findings] [5]. The respiratory effects of cholinergic poisoning tend to be dramatic and are considered to be the major factor, leading to the death of its victims.

Agents that cause these cholinergic clinical effects inhibit AChE, and include organophosphate and carbamate insecticides (cholinesterase inhibitors), certain species of mushrooms that contain muscarinic and nicotinic receptor agonists such as nicotine.

Anticholinergic toxidrome

Characteristics of the anticholinergic syndrome are essentially the opposite of the cholinergic syndromes, and have long been taught using the old medical adage, "dry as a bone, blind as a bat, red as a beet, hot as a hare, and mad as a hatter," which corresponds with a symptomatic person's anhidrosis, mydriasis, flushing, fever, and delirium, respectively. Depending on the dose and time following exposure, various CNS effects may manifest from an anticholinergic agent. Restlessness, apprehension, abnormal speech, confusion, agitation, tremor, picking movements, ataxia, stupor, and coma all have been described following exposure to various anticholinergics. Compounds exhibiting anticholinergic actions competitively block the action of acetylcholine at its receptors, both centrally and peripherally. Numerous substances can cause the anticholinergic syndrome. Some of the more common agents include antihistamines, atropine, cyclic antidepressant drugs, phenothiazines, anti-Parkinson's drugs, cyclobenzaprine, scopolamine, and several plants such as *Datura stramonium* (Jimson weed).

Opioid toxidrome

Opioids can induce coma, respiratory depression, bradycardia, hypotension, hypothermia, miosis, pulmonary edema, decreased bowel sounds, and decreased reflexes. Common causes of this syndrome, in which CNS depression and miosis are two key signs, include morphine, codeine, diacetylmorphine (heroin), oxycodone, hydrocodone, hydromorphone, and methadone. Meperidine and propoxyphene toxicity has been associated with mydriasis (pupil dilation), but not with miosis (pupil constriction).

TABLE 52.1 Summary of the toxidromes.

Toxidrome	Signs and symptoms	Common toxins
Anticholinergic	Agitation	Antihistamines, anti-Parkinson's drugs, antipsychotic drugs, and tricyclic antidepressants
	Blurred vision	
	Decreased bowel sounds	
	Dry skin	
	Fever	
	Flushing	
	Hallucinations	
	Ileus	
	Lethargy/coma	
	Mydriasis	
	Myoclonus	
	Psychosis	
	Seizures	
	Tachycardia	
	Urinary retention	
Cholinergic	Diarrhea	Organophosphate and carbamate insecticides, and some mushrooms
	Urination	
	Miosis	
	Bradycardia	
	Bronchorrhea	
	Emesis	
	Lacrimation	
	Salivation	
Opioid	Bradycardia	Opiates (e.g., morphine and codeine), propoxyphene, and alpha-2 agonists (e.g., clonidine and oxymetazoline).
	Decreased bowel sounds	
	Hypotension	
	Hypothermia	
	Lethargy/coma	
	Miosis	
	Shallow respirations	
	Slow respiratory rate	

(*Continued*)

TABLE 52.1 (Continued)

Toxidrome	Signs and symptoms	Common toxins
Sedative-hypnotic	Ataxia	Benzodiazepines, barbiturates, ethanol, and meprobamate
	Blurred vision	
	Confusion	
	Diplopia	
	Dysesthesia	
	Hypotension	
	Lethargy/coma	
	Nystagmus	
	Respiratory depression	
	Sedation	
	Slurred speech	
Sympathomimetic	Agitation	Amphetamines, caffeine, cocaine, ephedrine, and phencyclidine
	Diaphoresis	
	Excessive motor activity	
	Excessive speech	
	Hallucinations	
	Hypertension	
	Hyperthermia	
	Insomnia	
	Restlessness	
	Tachycardia	
	Tremor	

Numerous drugs can also mimic the opioid syndrome by inducing coma, respiratory depression, and miosis, including alpha-2 agonists (e.g., clonidine and oxymetazoline) and antipsychotics.

Sedative hypnotic toxidrome

Sedative-hypnotics are a broad class of drugs (e.g., benzodiazepines, barbiturates, meprobamate, ethchlorvynol, and zolpidem) that can induce sedation, respiratory depression, hypotension, hyporeflexia, nystagmus, dysarthria, staggering gait, apnea, and coma. Numerous sedative-hypnotics have been utilized to commit drug-facilitated crimes (see below), such as the solution of chloral hydrate in ethanol. There are few diagnostic clinical features that distinguish the drugs in this class from one another.

Sympathomimetic toxidrome

Their physiological and toxic effects associated with the sympathomimetic toxidrome reflect increased sympathetic activity due to increased release of catecholamines, inhibition of neurotransmitter reuptake, and/or direct stimulation of alpha- and beta-adrenergic receptors. Physiologic responses to activation of the adrenergic system are complex and depend on the type of receptor activated; some are excitatory and others have opposing inhibitory responses. Stimulation of the sympathetic nervous system produces CNS excitation (agitation, anxiety, tremors, delusions, and paranoia), tachycardia, hypertension, mydriasis, hyperpyrexia, and diaphoresis. In severe cases, seizures, cardiac arrhythmias, and coma may occur. Examples of drugs that produce a sympathomimetic

response include amphetamines, COC, phencyclidine (PCP), ephedrine, methcathinone, and pseudoephedrine.

Key laboratory formulas

Anion gap

Obtaining a basic metabolic panel in all poisoned patients is recommended and is an important initial screening test. When low serum bicarbonate is discovered on a metabolic panel, the clinician should determine if an anion gap (AG) exists. The formula most commonly used for the AG calculation is as follows:

$$\mathrm{AG} = \left[\mathrm{Na}^{+}\right] - \left[\mathrm{Cl}^{-} + \mathrm{HCO}_3^{-}\right].$$

The primary cation (sodium) and anions (chloride and bicarbonate) are represented in the equation. Other serum cations are not commonly included in this calculation, because their concentrations are relatively low (e.g., potassium) or assigning a number to represent their respective contribution is difficult (e.g., magnesium and calcium). Similarly, a multitude of other serum anions (e.g., sulfate, phosphate, and organic anions) are not commonly measured and therefore are excluded when determining the AG. These "unmeasured" ions represent the gap calculated using the previous equation. The normal range for an AG is typically to be 8−16 mmol/L, but it has been suggested that, because of changes in the technique used to measure chloride, some have suggested a reference interval of 6−14 mmol/L [6]. Practically speaking, useful mnemonics for causes of high AG metabolic acidoses are the classic MUDPILES (representing methanol, uremia, diabetes, paraldehyde, iron (and isoniazid), lactate, ethylene glycol, and salicylate) and the more recently proposed GOLD MARK (glycols (ethylene and propylene), oxoproline, L-lactate, D-lactate, methanol, aspirin, renal failure, and ketoacidosis; see [7]).

Osmolal gap

The main osmotically active constituents of serum are Na^{+}, Cl^{-}, HCO_3^{-}, glucose, and urea. Several formulas based on measurement of these substances have been used to estimate serum osmolality [7]. In practice, one has not shown itself to be superior to the others, yet each equation demonstrates significant differences in the osmolal gap (OSMg) reference interval [8]. Therefore reference intervals must be validated on appropriate patient populations. Two commonly used formulas (in conventional and SI units) are presented here:

$$\mathrm{OSMc}\left(\mathrm{mOsm/kg}\right) = 2\,\mathrm{Na}\left(\mathrm{mmol/L}\right) + \mathrm{glucose}\left(\mathrm{mg/dL}\right)/18 + \mathrm{urea}\left(\mathrm{mg/dL}\right)/2.8$$

$$\mathrm{OSMc}\left(\mathrm{mOsm/kg}\right) = 2\,\mathrm{Na}\left(\mathrm{mmol/L}\right) + \mathrm{glucose}\left(\mathrm{mmol/L}\right) + \mathrm{urea}\left(\mathrm{mmol/L}\right)$$

or

$$\mathrm{OSMc}\left(\mathrm{mOsm/kg}\right) = 1.86\,\mathrm{Na}\left(\mathrm{mmol/L}\right) + \mathrm{glucose}\left(\mathrm{mg/dL}\right)/18 + \mathrm{urea}\left(\mathrm{mg/dL}\right)/2.8 + 9$$

$$\mathrm{OSMc}\left(\mathrm{mOsm/kg}\right) = 1.86\,\mathrm{Na}\left(\mathrm{mmol/L}\right) + \mathrm{glucose}\left(\mathrm{mmol/L}\right) + \mathrm{urea}\left(\mathrm{mmol/}L\right) + 9.$$

The difference between the measured osmolality (OSMm), measured by freezing-point depression, and the calculated osmolality (OSMc) is referred to as delta-osmolality, or the OSMg

$$\mathrm{OSMg} = \mathrm{OSMm} - \mathrm{OSMc}.$$

Traditionally, a normal OSMg has been defined as 10 mOsm/kg or less. Researchers have found the OSMg to vary depending on the population studied. Elevated OSMg implies the presence of unmeasured osmotically active substances. Volatile alcohols (ethanol, methanol, isopropanol, acetone, and ethylene glycol), when present at significant concentrations, may cause an increased serum osmolality and a subsequent increase in the OSMg. However, it is important to remember that volatile alcohols are not detected when osmolality is measured with a vapor pressure osmometer. Therefore, for the purpose of determining the OSMg, only osmolality measurements based on freezing-point depression are acceptable.

A significant residual OSMg (>10 mOsm/kg) would suggest the possible presence of isopropanol, methanol, acetone, or ethylene glycol. This information, in conjunction with the presence or absence of metabolic acidosis or serum acetone, is helpful to the clinician when specific measurements of alcohols other than ethanol and of ethylene glycol are not available on an emergency basis (Table 52.2). It must be realized that ketones and substances administered to patients, such as polyethylene glycol (burn cream), mannitol (osmotic diuretic), and propylene glycol (solvent for medications including diazepam and phenytoin), may also increase serum osmolality.

Pharmacology and analysis of specific drugs and toxic agents

The toxic, pharmacologic, biochemical, and analytical characteristics of several individual drugs and toxins are discussed in this section.

Agents that cause cellular hypoxia

CO and methemoglobin-forming agents interfere with oxygen transport, resulting in cellular hypoxia. Cyanide

TABLE 52.2 Laboratory findings characteristic of ingestion of alcohols.

Alcohol	Serum osmolal gap	Metabolic acidosis with anion gap	Serum acetone	Urine oxalate
Ethanol	+	−	−	−
Methanol	+	+	−	−
Isopropanol	+	−	+	−
Ethylene glycol	+	+	−	+

+, Present; −, absent.

interferes with oxygen use and therefore causes an apparent cellular hypoxia.

Carbon monoxide

CO is a colorless, odorless, and tasteless gas that is a product of incomplete combustion of carbonaceous material. Common exogenous sources of carbon monoxide (CO) include cigarette smoke, gasoline engines, and improperly ventilated home heating units. Small amounts of CO are produced endogenously in the metabolic conversion of heme to biliverdin.

CO complexes tightly with the heme Fe^{2+} of hemoglobin to form carboxyhemoglobin. The binding affinity of hemoglobin for CO is ∼250 times greater than that for oxygen. Thus high concentrations of carboxyhemoglobin limit the oxygen content of blood. Moreover, the binding of CO to a hemoglobin subunit increases the oxygen affinity for the remaining subunits in the hemoglobin tetramer. At a given tissue PO_2 value, less oxygen dissociates from hemoglobin when CO is also bound, thereby shifting the hemoglobin−oxygen dissociation curve to the left. Consequently, CO not only decreases the oxygen content of blood, but also decreases oxygen availability to tissue, thereby producing a greater degree of tissue hypoxia than would result from an equivalent reduction in oxyhemoglobin due to hypoxia alone [9].

The toxic effects of CO are a result of tissue hypoxia. Organs with high oxygen demand, such as the heart and brain, are more sensitive to hypoxia and thus account for the major clinical sequelae of CO poisoning. It must be emphasized that the carboxyhemoglobin concentration, although helpful in diagnosing CO exposure, does not always correlate with the clinical findings or prognosis [7]. Other factors that contribute to toxicity include length of exposure, metabolic activity, and underlying disease, especially cardiac or cerebrovascular disease. Treatment for CO poisoning involves removal of the individual from the contaminated area and administration of oxygen. The half-life ($t_{1/2}$) of carboxyhemoglobin in the body is variable; in room air, the approximate $t_{1/2}$ is ∼4−5 hours; during hyperbaric oxygen therapy, it is as short as 12−20 minutes [10].

CO may be released from hemoglobin and then measured by GC, or it may be determined as carboxyhemoglobin by spectrophotometry. GC methods are extremely accurate and precise for very low concentrations of CO. However, spectrophotometric methods are more widely used, since they are rapid, convenient, accurate, and precise, except at very low concentrations of carboxyhemoglobin (<2%−3%).

Methemoglobin

An acquired methemoglobinemia may be caused by various drugs and chemicals, as well as by oxides of nitrogen and other oxidant combustion products; consequently, prolonged smoke inhalation may be a potential cause of methemoglobinemia [7]. The heme iron in hemoglobin is normally present in the ferrous state (Fe^{2+}). When oxidized to the ferric state (Fe^{3+}), methemoglobin is formed, and this form of hemoglobin cannot bind oxygen. The normal percentage of methemoglobin is <1.5% of total hemoglobin. The severity of symptoms usually correlates with measured methemoglobin levels; methemoglobin percentages up to 20% may cause slate-gray cutaneous discoloration, cyanosis, and chocolate-brown blood. Percentages between 20% and 50% may cause dyspnea, exercise intolerance, fatigue, weakness, and syncope. More severe symptoms of dysrhythmias, seizures, metabolic acidosis, and coma are associated with methemoglobin percentages of 50%−70%, and >70% may be lethal [11].

All of these symptoms are a consequence of hypoxia that is associated with the diminished O_2 content in the blood, and with decreased O_2 dissociation from hemoglobin species in which some, but not all, subunits contain heme iron in the ferric state. The PO_2 is normal in these patients, and therefore is the calculated hemoglobin oxygen saturation. Thus a normal PO_2 in a cyanotic patient is a significant indication for the possible presence of methemoglobinemia.

Specific therapy for toxic methemoglobinemia involves the administration of methylene blue, which acts as an electron transfer agent in the NADPH−methemoglobin reductase reaction. Ascorbic acid can also reverse methemoglobin by an alternate metabolic pathway, but is of minimal use acutely because of its slow action [12].

Direct measurement of methemoglobin may be performed by the spectrophotometric method of Evelyn and Malloy [13] or by automated multiwavelength measurements with a cooximeter.

Cyanide

Cyanide salts are used in numerous industries, such as metallurgy, photographic developing, plastic manufacturing, fumigation, and mining. Other sources include organic compounds acetonitrile, plants, and drugs such, as nitroprusside, which contains five cyanide molecules. Cyanide readily crosses all biological membranes, avidly binds to heme iron (Fe^{3+}) in the cytochrome $a-a_3$ complex within mitochondria, and causes decoupling of oxidative phosphorylation [14]. Patients exposed to toxic concentrations of cyanide may exhibit rapid onset of symptoms typical of cellular hypoxia, including the following: flushing; headache; nausea and vomiting; anxiety, confusion, and collapse; initial hypertension and tachycardia progressing to hypotension; cyanosis bradycardia; apnea; coma; seizures; complete heart block; and death if the dose is sufficiently large [15]. If cyanide toxicity is suspected, a cyanide antidote may be administered without waiting for laboratory confirmation. Available antidotes are hydroxocobalamin (Cyanokit) and sodium thiosulfate and sodium nitrite (Nithiodote).

Cyanide measurements can be performed using several methodologies, including spectrometry following microdiffusion separation, GC with electron capture, nitrogen selective or MS detection methods, or LC followed by fluorescence and MS detection methods [16].

Alcohols

Several alcohols are toxic and medically important; they include ethanol, methanol, isopropanol, acetone (also a metabolite of isopropanol), and ethylene glycol.

Ethanol

Ethanol is the most widely used and often abused chemical substance. Consequently, measurement of ethanol is one of the most frequently performed tests in the toxicology laboratory. Ethanol is considered a CNS depressant. Symptoms vary from euphoria and decreased inhibitions, to increased disorientation and incoordination, and then to coma and death (Table 52.3). When consumed with other CNS depressant drugs, ethanol exerts a potentiation or synergistic depressant effect. This can occur at relatively low alcohol concentrations, and numerous deaths have resulted from combined ethanol and drug ingestion [17]. Ethanol is metabolized principally by liver alcohol dehydrogenase to acetaldehyde, which is subsequently oxidized to acetic acid by aldehyde dehydrogenase. The rate of elimination of ethanol from blood approximates a zero-order process. This rate varies among individuals, averaging ~15 mg/dL/h for males and 18 mg/dL/h for females [18].

TABLE 52.3 Blood alcohol concentrations and clinical effects.

Blood alcohol concentrations (mg/dL)	Clinical effects
50−100	Sedation and increased reaction times
100−200	Impaired motor function, slurred speech, and ataxia
200−300	Emesis and stupor
300−400	Coma
>500	Respiratory depression and death

Ethanol is also a teratogen, and alcohol consumption during pregnancy can result in a baby with fetal alcohol spectrum disorder. These effects may include physical, mental, behavioral, and/or learning disabilities with possible lifelong implications and are 100% preventable when a woman completely abstains from alcohol during her pregnancy.

Methanol

Methanol is used as a solvent in several commercial products, including windshield wiper fluid, copy machine fluids, fuel additives, paint remover or thinner, antifreeze, canned heating sources, deicing fluid, shellacs, and varnishes. It may be consumed intentionally by individuals as an ethanol substitute, for suicide, or accidentally when present as a contaminant in illegal whiskey, or inadvertently by children.

The CNS effects of methanol are substantially less severe than those of ethanol. Methanol is oxidized by liver alcohol dehydrogenase (at about one-tenth the rate of ethanol) to formaldehyde. Formaldehyde, in turn, is rapidly oxidized by aldehyde dehydrogenase to formic acid, which may cause serious acidosis, optic neuropathy, potentially resulting in blindness, CNS bleeding, or death [7]. The mainstay of therapy for methanol toxicity includes the administration of fomepizole or ethanol as a

competitive alcohol dehydrogenase inhibitor, and hemodialysis can be performed when required. Methanol poisoning can be lethal if not recognized early. Unfortunately, in some instances, a latent period can be as long as 12–24 hours before toxicity is recognized, making laboratory identification of this poisoning critical [19].

Isopropanol and acetone

Isopropanol is readily available as a 70% aqueous solution for use as rubbing alcohol, but can also be found in cleaners, disinfectants, antifreezes, cosmetics, solvents, inks, and pharmaceuticals. It has about twice the CNS depressant effect as ethanol [16]. Isopropanol is a secondary alcohol (the hydroxyl group is attached to a central, rather than a terminal carbon), and it is metabolized by alcohol dehydrogenase to a ketone, not an acid; Therefore ingestions do not cause significant metabolic acidosis [20]. Isopropanol has a short $t_{1/2}$ of 2.5–3.0 hours, as it is rapidly metabolized to acetone, which is eliminated much more slowly ($t_{1/2}$, 3–6 hours) [16]. Therefore concentrations of acetone in serum often exceed those of isopropanol during the elimination phase following isopropanol ingestion. Acetone has CNS depressant activity similar to that of ethanol, and, because of its longer $t_{1/2}$, it may prolong the apparent CNS effects of isopropanol [16]. Supportive care is the mainstay of treatment, with rare reports of dialysis in cases of severe intoxication.

Ethylene glycol

Ethylene glycol is present in antifreeze products, deicing products, detergents, paints, and cosmetics. It may be ingested accidentally or for the purpose of inebriation or suicide. Because it tastes sweet, some animals are attracted to it. Veterinarians are often familiar with ethylene glycol toxicity because of cases involving dogs or cats that drank radiator fluid.

Ethylene glycol itself has initial CNS effects resembling those of ethanol [21]. However, metabolism of ethylene glycol by alcohol dehydrogenase to glycolaldehyde is then metabolized to glycolic, glyoxylic, and oxalic acids [16]. Oxalate readily precipitates with calcium to form insoluble calcium oxalate crystals. There is a lack of correlation between ethylene glycol concentration and severity of toxicity. It is thus impossible to define a serum ethylene glycol concentration associated with a high probability of death largely due to the timing of the blood collection in relation to the ingestion and the amount that had already been metabolized to more toxic metabolites. However, the serum glycolic acid concentration correlates more closely with clinical symptoms and mortality than does the concentration of ethylene glycol. Because of the rapid elimination of ethylene glycol ($t_{1/2}$, 2–5 hours), its serum concentration may be low or undetectable at a time when glycolic acid remains elevated [16]. Thus the determination of ethylene glycol and glycolic acid provides useful clinical and confirmatory analytical information in cases of ethylene glycol ingestion. The mainstay of therapy for ethylene glycol toxicity includes administration of ethanol or fomepizole as a competitive alcohol dehydrogenase inhibitor and dialysis.

Analysis of ethanol

Serum/plasma and blood ethanol

Serum, plasma, and whole blood are suitable specimens for the determination of ethanol. Alcohol distributes into the aqueous compartments of blood; because the water content of serum is greater than that of whole blood, higher alcohol concentrations are obtained with serum as compared with whole blood. Experimentally, the serum-to-whole blood ethanol ratio is 1.18 (1.10–1.35) [22] and varies slightly with hematocrit [23]. Therefore laboratories that perform alcohol determinations should make the specimen of choice clear. For legal purposes, whole blood is what the court system uses to determine if someone is driving under the influence. Therefore, if a serum or plasma alcohol was obtained, it must be converted into its whole blood equivalent when used at trial.

To measure ethanol in serum/plasma, enzymatic analysis is the method of choice for many laboratories. In this method, ethanol is measured by oxidation to acetaldehyde with NAD, a reaction catalyzed by ADH. With this reaction, the formation of NADH, measured at 340 nm, is proportional to the amount of ethanol in the specimen [24]:

$$\text{Ethanol} + \text{NAD} \xrightarrow{\text{ADH}} \text{Acetaldehyde} + \text{NADH}.$$

Under most assay conditions, ADH is reasonably specific for ethanol, with interferences by isopropanol, acetone, methanol, and ethylene glycol of typically $<1\%$. As a precaution, the venipuncture site is recommended to be cleansed with an alcohol-free disinfectant, such as aqueous benzalkonium chloride. Since NADH is responsible for the measurable complex leading to alcohol findings with the enzymatic test for ethanol, it is known that, for some assays when patients have significantly elevated lactate and lactate dehydrogenase, it can result in a spurious positive finding, and it is suspected that other dehydrogenases and substrates may cause similar interference [25].

Analysis of volatile alcohols (methanol, isopropanol, and acetone)

Development of GC methods for volatiles plays a significant step in the recognition and treatment of volatile alcohols. Flame ionization GC remains the most common

method for the detection and quantitation of volatile alcohols in biological samples [16]. Not only does it distinguish between ethanol, methanol, isopropanol, and acetone, but also it has the capability to measure concentrations as low as 10 mg/dL (0.01%). Headspace injection GC is an excellent method for the measurement of methanol, isopropanol, acetone, and ethanol. In addition, an adaptation of this technique may be used to measure formate, the toxic metabolite of methanol, after esterification to methyl formate. Conversely, direct injection GC is the method of choice for ethylene glycol, because it has a higher boiling point and is not as amenable to headspace analysis.

Ethanol biomarkers

Ethyl glucuronide (EtG), ethyl sulfate (EtS), and phosphatidylethanols (PEths) are biomarkers of ethanol consumption. EtG and EtS are phase II metabolites of ethanol. EtG is formed through the UDP-glucuronosyltransferase catalyzed conjugation of ethanol with glucuronic acid [26], and EtS is also formed directly by the conjugation of ethanol with sulfate group [27]. PEths are formed by phospholipase D in the presence of ethanol [28]. Due to the long urinary elimination time ($\leq$80 hours), its specificity for ethanol exposure, and its low detection limits, EtG is utilized as a marker of recent ethanol intake in a variety of clinical settings, including medical monitoring for relapse, emergency department patient evaluation, and transportation accident investigation.

Monitoring both EtG and EtS in urine improves sensitivity and specificity. EtG, but not EtS, can be either produced or decreased postspecimen collection [29]. Upper respiratory infections, as well as beta-glucuronidase hydrolysis, may lower levels of EtG but do not impact EtS [30]. However, there are challenges associated with the implementation of EtG and EtS to assess prior ethanol consumption. These include the absence of cutoff concentrations to distinguish between drinking and incidental exposure, such as nonbeverage sources of ethanol exposure (i.e., hand sanitizers and mouthwash) [31,32], nonstandardization of laboratory reporting limits, potential sample stability limitations, and microbial activity in collected samples. All of these factors may complicate interpretation of results [33]. However, some interpretive guidance recommendations have been proposed; EtG of $>$1000 ng/mL may indicate: heavy drinking in the previous 48 hours, or light drinking the same day. An EtG from 500 to 1000 ng/mL may indicate: heavy drinking in the previous 3 days, light drinking in the past 24 hours, or intense "extraneous exposure" within 24 hours. Values of $<$500 ng/mL may indicate: heavy drinking in the previous 3 days, light drinking in the past 36 hours, or recent "extraneous" exposure [30].

PEths are a group of phospholipids with a common phosphoethanol head group with two fatty acid chains that differ in chain length and degree of unsaturation [28]. PEth is a promising marker because of its persistence in blood for as long as 3 weeks after only a few days of moderately heavy drinking (about four drinks per day) [30].

Analgesics (nonprescription)

Analgesics are substances that relieve pain without causing loss of consciousness. When used in excess, analgesics such as acetaminophen and salicylate can result in a toxic response.

Acetaminophen

Acetaminophen (*N*-acetyl-*p*-aminophenol, paracetamol) has analgesic and antipyretic actions. In common with the group of drugs referred to as nonsteroidal antiinflammatory drugs (e.g., aspirin, ibuprofen, and indomethacin), the pharmacologic actions of acetaminophen are related to its competitive inhibition of cyclooxygenase enzymes. This results in decreased production of prostaglandins, which are important mediators of inflammation, pain, and fever [34]. In normal doses, acetaminophen is safe and effective, but may cause severe hepatotoxicity or death when consumed in overdose quantities.

Acetaminophen is normally metabolized in the liver to glucuronide (50%−60%) and sulfate ($\approx$30%) conjugates (Fig. 52.1). A smaller amount ($\approx$10%) is metabolized by a cytochrome P450 mixed-function oxidase pathway that is thought to form the highly reactive intermediate, *N*-acetyl-*p*-benzoquinone imine. This intermediate normally undergoes electrophilic conjugation with glutathione and then subsequent transformation to cysteine and mercapturic acid conjugates of acetaminophen. With acetaminophen overdose, the sulfation pathway becomes saturated; consequently, a greater portion is metabolized by the

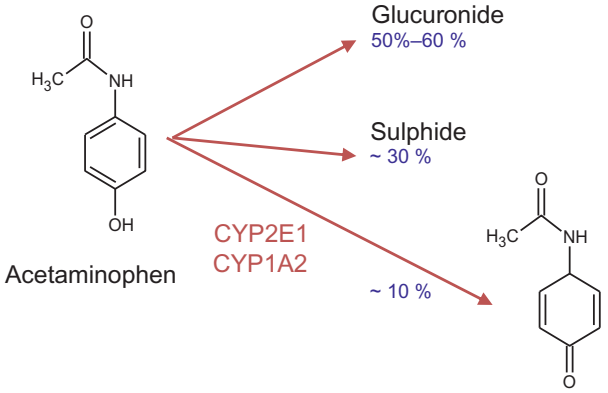


FIGURE 52.1 Metabolic pathway of acetaminophen.

P450 mixed-function oxidase pathway. When the tissue stores of glutathione become depleted, arylation of cellular molecules by the benzoquinoneimine intermediate leads to hepatic necrosis [35].

The initial clinical findings in acetaminophen toxicity may be absent of relatively mild and nonspecific (nausea, vomiting, and abdominal discomfort); hepatic necrosis typically begins 24−36 hours after toxic ingestion and becomes most severe by 72−96 hours [36]. Although uncommon with severe overdose, coma and metabolic acidosis may occur before development of hepatic necrosis [37]. Specific therapy for acetaminophen overdose is the administration of *N*-acetylcysteine (NAC), which acts as a glutathione substitute and provides substrate to replenish hepatic glutathione or to enhance sulfate conjugation, or both. The time of administration of NAC is critical. Maximum efficacy is observed when NAC is administered within 8 hours, and efficacy declines with time; therefore it is most effective when administered before hepatic injury occurs, as signified by elevations of AST and ALT [7].

The measurement of serum acetaminophen concentrations becomes important for proper assessment of the severity of overdose and for appropriate decision making for antidotal therapy. If serum acetaminophen results are not available locally, treatment with NAC should begin until levels are available. The Rumack−Matthew nomogram relates serum acetaminophen concentration and time following acute ingestion to the probability of hepatic necrosis [38]. However, the Rumack−Matthew nomogram suffers from many limitations including it begins at 4-hour postingestion, it was developed in adult patients only after a single acute ingestion, and the time of ingestion must be known. Foreknowledge of all of these parameters are not frequently known in many instances of acetaminophen overdose.

Many spectrophotometric methods are available for the determination of acetaminophen. In general, these methods are relatively easy to perform but are subject to significant interferences, such as bilirubin or bilirubin byproducts, absorbing at similar wavelengths, and thus may produce especially misleading results [7]. Methods specific for parent acetaminophen should be used to prevent interferences with nontoxic metabolites. Immunoassays are widely used for this purpose, as they are rapid, easily performed, and accurate. Most chromatographic methods are very accurate and are considered reference procedures. A qualitative, one-step lateral flow immunoassay (cutoff of 25 μg/mL) may be used for point-of-care testing, but it has a low positive predictive value [39].

Salicylate

Acetylsalicylic acid (aspirin) has analgesic, antipyretic, and antiinflammatory properties. These therapeutic benefits derive from its ability to inhibit biosynthesis of prostaglandins by inhibition of cyclooxygenase enzymes (COX-1 and COX-2 isoenzymes) [34]. Aspirin also interferes with platelet cyclooxygenase, reducing the formation of thromboxane A_2, a potent mediator of platelet aggregation.

Salicylate is eliminated mainly by conjugation with glycine and to a lesser extent with glucuronic acid [34]. These metabolic pathways may become saturated even at high therapeutic doses. Consequently, serum salicylate concentrations may increase disproportionately with dosage, and there is a relationship between dose and salicylate clearance. Salicylate elimination $t_{1/2}$ is prolonged at higher doses (15−30 hours) as compared with lower doses (2−3 hours). Following acute salicylate overdose, patients may initially be asymptomatic, especially if the product is enteric-coated, which delays its dissolution in the gastric environment. Salicylate toxic patients may develop nausea, vomiting, abdominal pain, tinnitus, tachypnea, oliguria, and altered mental status, ranging from agitation to lethargy to coma. Salicylates directly stimulate the central respiratory center and thereby cause hyperventilation and respiratory alkalosis. In addition, salicylates cause uncoupling of oxidative phosphorylation. As a result, heat production (hyperthermia), oxygen consumption, and metabolic rate may be increased. Salicylates also enhance anaerobic glycolysis but inhibit the Krebs cycle and transaminase enzymes, which leads to accumulation of organic acids and thus to metabolic acidosis. The primary acid-base disturbance observed with salicylate overdosage depends on age and severity of intoxication [7].

Toxic doses of aspirin may lead to the formation of concretions or bezoars in GI tract and pylorospasm, which can lead to delays in drug absorption. Serum salicylate in such instances may not reach maximum concentrations for 6 hours or longer. Concentrations alone are of poor prognostic value; however, certain clinical findings are predictors of a poor prognosis, including pulmonary edema, fever, coma, and acidosis. Initial serial concentrations should be performed every 2 hours while the patient is monitored clinically. Interpretation of salicylate concentrations as a guide for clinical management decisions can be difficult. Perhaps, the most well-known attempt at utilizing salicylate concentrations to predict the severity of salicylate toxicity was the nomogram developed by Done [40]. However, the Done nomogram was primarily based on healthy pediatric patients with a single acute ingestion of nonenteric-coated salicylate. As a result, it is often regarded as having limited clinical utility by providers. Treatment for salicylate intoxication is directed toward: (1) decreasing further absorption; (2) increasing elimination; and (3) correcting acid-base and electrolyte disturbances. Activated charcoal binds aspirin and

prevents its absorption. Elimination of salicylate may be enhanced by alkaline diuresis and in severe cases by hemodialysis. Sodium bicarbonate may be given to alleviate metabolic acidosis [7].

Classic methods for the measurement of salicylate in serum are based on the Trinder method [41]. These procedures rely on the reaction between salicylate and Fe^{3+} to form a colored complex that is measured at 540 nm. Other methods for salicylate quantitation include fluorescent polarization immunoassay and a salicylate hydroxylase-mediated photometric technique [7]. These procedures are subject to some of the same interferences as the Trinder method, but the salicylate hydroxylase method is considered more specific and has been adapted to automated analyzers. GC and LC methods are the most specific methods for salicylate [16], but their general availability, especially for emergency use, is limited.

Drugs of abuse

Drug use and abuse are widespread in society, and public awareness has been heightened as to their impact on public safety and on lost productivity in the workplace. To resolve these issues, governmental, industrial, educational, and sport agencies are increasingly requiring drug testing of prospective and existing employees, students, and participants in professional and amateur athletics. Moreover, drug abuse during pregnancy is a matter of concern, both medically and socially [42]. Testing for drugs of abuse may be a medical requirement for (1) organ transplantation candidates; (2) pain management clinics; (3) drug abuse treatment programs; and (4) psychiatric programs. Drug testing for these purposes represents a significant activity for toxicology laboratories. Although there are many drugs and drug classes that are abused, and there is geographic variation in their identity and prevalence, not all can be discussed in this chapter; therefore only the ones that are the most commonly encountered will be discussed.

Testing for drugs of abuse usually involves testing a single urine specimen for a variety of drugs. It should be noted, however, that testing of a single random urine drug specimen detects only recent drug use (Table 52.4), and it does not differentiate casual use from chronic drug abuse. Urine drug testing (UDT) also cannot determine the degree of impairment, the dose of drug taken, or the exact time of use.

Amphetamine-type stimulants

Several stimulants and hallucinogens chemically related to phenylethylamine are referred to collectively as amphetamine-type stimulants (ATSs). They are considered to be sympathomimetic drugs, meaning that they mimic endogenous transmitters in the sympathetic nervous system [43].

Amphetamine and methamphetamine

Amphetamine and methamphetamine are CNS stimulant drugs that have some legitimate pharmacologic use [44], including narcolepsy, obesity, and attention-deficit hyperactivity disorders (ADHDs). They produce an initial euphoria and have a high abuse potential. These drugs have a stimulating effect on both the central and peripheral nervous systems. In the brain, a primary action is to elevate the concentrations of extracellular monoamine neurotransmitters (dopamine, serotonin, and norepinephrine) by promoting presynaptic release from the nerve endings [45]. Amphetamine and methamphetamine are

TABLE 52.4 Common classes of medications and detection window in urine.

Drug/drug class	Average detection window (urine)
Amphetamine/methamphetamine	~3 days
Barbiturates	Short-acting: 4–6 days
	Long-acting: 2–3 weeks
Benzodiazepines	~1–10 days
Cocaine and metabolite (benzoylecgonine)	~3 days
THC	3–90 days depending on usage
Lysergic acid diethylamide	~10 days
Opioids/opiates	~3 days
Phencyclidine	~8 days

THC, Tetrahydrocannabinol.

substrates for the dopamine, serotonin, and norepinephrine transporters.

Amphetamine and methamphetamine: (1) increase blood pressure, heart rate, body temperature, and motor activity; (2) relax bronchial muscle; and (3) depress the appetite. Abuse of these drugs may lead to strong psychological dependence, marked tolerance, mild physical dependence, and a severe form of chronic intoxication psychosis similar to schizophrenia. Tolerance and psychological dependence develop with repeated use of amphetamines [44].

Amphetamine is extensively metabolized to a variety of metabolites, including norephedrine and *p*-hydroxyamphetamine, both of which are pharmacologically active, and may be glucuronidated prior to excretion [43]. Methamphetamine is metabolized in liver primarily by hydroxylation and, to a lesser extent, by *N*-demethylation to amphetamine. Renal elimination is dependent on urine pH, and, although typically ~30% of a dose is excreted unchanged, this may vary from as much as 74% in acidic urine to as little as 1% in alkaline urine [16]. Therefore elimination $t_{1/2}$ (renal excretion and hepatic metabolism) varies from 7 to 34 hours depending on urine pH [46]. Similarly, methamphetamine is eliminated in urine in a pH-dependent manner, similar to amphetamine.

Designer stimulants

The terms "designer drugs" and "club drugs" were originated in the 1980s [47]. These drugs include phenylethylamine, benzylpiperazine, phenylpiperazine, pyrrolidinophenone, and cathinone and derivatives, and have gained popularity and notoriety among people who participate in all-night dance parties (raves) and those who visit nightclubs. Most designer drugs produce feelings of euphoria and energy and a desire to socialize; they also promote social and physical interactions. They are used at these events to enhance energy for prolonged partying and/or dancing, and to distort or enhance visual and auditory sensations. Due to the large variety of designer amphetamines, only a few will be discussed here.

3,4-Methylenedioxymethamphetamine and 3,4-methylenedioxyamphetamine

Methylenedioxymethamphetamine (MDMA, also known as "ecstasy" and "molly") is categorized as a stimulant as a result of its sympathomimetic effects, but is also categorized as an empathogen–entactogen [47]. Similar to amphetamine and methamphetamine, MDMA causes release of biogenic amines by reversing the action of their respective transporters. It has a preferential affinity for the serotonin transporter and therefore most strongly increases the extracellular concentration of serotonin [45]. MDMA undergoes hepatic metabolism including *N*-demethylation to the active drug methylenedioxyamphetamine (MDA) [48], paramethoxyamphetamine (PMA), and paramethoxymethamphetamine (PMMA): PMA and PMMA are methoxylated phenylethylamine derivatives with effects similar to but more potent than those of MDMA [49]. PMA is a metabolite of PMMA [50], but it is also an especially toxic designer amphetamine that has resulted in several deaths from its unsuspected ingestion as an ecstasy substitute. PMA is 10 times more active than MDMA in elevating brain serotonin concentrations and inhibiting serotonin uptake, but it has only a few effects on the dopamine system [51]. Multiple deaths have been associated with its use; symptoms usually mimic serotonin syndrome and include hyperthermia, tachycardia, seizures, cardiac dysrhythmias, and coma.

Cathinone and derivatives

Cathinones are β-ketoamphetamines with structural similarity to dopamine and amphetamine. Cathinone is a naturally occurring stimulant found in the leaves of the Khat plant (*Catha edulis*). Synthetic derivatives of cathinone emerged as drugs of abuse in approximately 2003 and were sold as "Bath Salts" and labeled as "not for human consumption." As new designer cathinone derivatives are produced by clandestine laboratories, the US government attempts to schedule as many of these drugs as possible (Synthetic Drug Abuse Prevention Act of 2012 and Synthetic Cathinones Control Act of 2013). The race continues to schedule synthetic drugs with high abuse potential and no known medicinal value.

Cathinone derivatives are abused because of their psychostimulant and hallucinatory effects, but have also been involved with human toxicity and fatality. Cathinones block dopamine and serotonin neurotransmitter reuptake, thereby prolonging the effects at peripheral and central neural junctions [52]. Common clinical presentations in patients are agitation, paranoia, auditory and tactile hallucinations, psychosis, myoclonus, and headaches. These symptoms can be severe and last for several days. Peripheral effects may include hyperthermia, hypertension, tachycardia, hyponatremia, nausea, vomiting, and chest pain as seen during sympathomimetic toxidrome.

Methylphenidate (Ritalin)

Methylphenidate (MPH) is a phenethylamine derivative with psychostimulant properties similar to amphetamine but is chemically distinct. It is commonly used to treat ADHD and narcolepsy [45]. Its pharmacologic properties are essentially the same as those of the amphetamines. MPH is rapidly metabolized to ritalinic acid and has a $t_{1/2}$ of ~4 hours. Diversion and abuse of MPH have been increasing among children and adults because of its stimulant and purported aphrodisiac properties. In overdose,

the clinical effects of MPH are similar to those of amphetamine and produce signs of generalized CNS stimulation that may lead to convulsions [45].

Analytical methods for amphetamine-type stimulants

Most "amphetamine" immunoassays have been designed to detect amphetamine/methamphetamine; some have been designed to detect MDMA and MDA and others to more broadly capture the ATS group—all with varying cross-reactivities [53]. Not all amphetamine immunoassays were suitable for detection of the amphetamine-derived designer drugs and especially not for the cathinone- and piperazine-derived substances [7]. Alternatively, other chemically related compounds, such as pseudoephedrine and phentermine, have been shown to produce positive results. Immunoassays from different manufacturers can also have very different "interference" profiles. On the other hand, detection of methylphenidate by immunoassays is problematic, since it does not cross-react with most commercial amphetamine immunoassays. The detection of MPH is also made difficult by its generally low concentration, while its metabolite, ritalinic acid, is present in much higher concentrations; it is also unstable upon storage, even when frozen [16].

In the end, any positive immunoassay for amphetamine or methamphetamine should be confirmed. In addition, if other designer stimulants are suspected, a negative immunoassay screen cannot rule out the presence of these drugs. Fortunately, numerous GC- and LC-based methods for identification and quantitation of these drugs in biological samples have been put forth.

Barbiturates

The success of barbital in 1903 and phenobarbital in 1912 [54] spawned the synthesis and testing of numerous derivatives, of which approximately 50 were distributed commercially. Because of their low therapeutic index and high potential for abuse, they have been largely replaced by the much safer benzodiazepines.

Barbiturates continue to be, although much less frequently than in the past, subject to abuse. Due to their rapid onset and short duration of action, the short- to intermediate-acting barbiturates that are used as sedative-hypnotics (amobarbital, butabarbital, butalbital, pentobarbital, and secobarbital) are most commonly abused. The longer-acting barbiturates (mephobarbital and phenobarbital), used primarily for their anticonvulsant properties, are rarely abused. The detection period in urine following ingestion of barbiturates varies with different assays and depends on the pharmacologic properties of the drugs. Short- to intermediate-acting barbiturates generally may be detected for 1−4 days following use; long-acting barbiturates, such as phenobarbital, may be detected for several weeks after long-term use. Anesthetic doses of barbiturates, such as pentobarbital, are used to reduce intracranial pressure from cerebral edema associated with head trauma, surgery, or cerebral ischemia [7]. Therefore appropriate analytical methods are necessary to monitor serum pentobarbital concentrations in these circumstances. The barbiturates produce varying degrees of CNS depression, ranging from mild sedation to general anesthesia. With long-term administration of gradually increasing doses, tolerance to effects on mood, sedation, and hypnosis occurs more readily, and is greater than tolerance to anticonvulsant and lethal effects; thus, as tolerance increases, the therapeutic index decreases [54].

Numerous commercial immunoassays for barbiturates are available. Most use antibodies directed toward secobarbital, and, although the degree of cross-reactivity of other barbiturates varies with each assay, most have sufficient cross-reactivity to detect the major therapeutically used barbiturates [16,55]. Several confirmation methods for barbiturates have been described. These include both GC and HPLC and MS [16].

Benzodiazepines

Benzodiazepines are any of a group of compounds having a common molecular structure and acting similarly as CNS depressants. The term *benzodiazepine* refers to the portion of the structure composed of a benzene ring fused to a seven-membered diazepine ring and a phenyl ring attached to the 5-position of the diazepine ring [55]. As a class of drugs, benzodiazepines are among the most commonly prescribed drugs in the Western hemisphere because of their (1) efficacy; (2) safety; (3) low addiction potential; (4) minimal side effects; and (5) high public demand for sedative and anxiolytic agents. Long-term benzodiazepine use poses a risk for the development of dependence and abuse, particularly for those agents with the shortest $t_{1/2}$, the highest potency (alprazolam and triazolam), and the greatest lipophilicity (diazepam). Regular use will produce tolerance to most of the adverse effects of benzodiazepines [56]. Tolerance may take weeks or months to develop, although this will depend on the dose of drugs used, the frequency of administration, and the pharmacokinetic $t_{1/2}$ of the drug.

The benzodiazepines given by themselves or in combination with other drugs, particularly narcotic analgesics (opioids), are among the most widely abused drugs. Their ability to suppress or dampen withdrawal symptoms and to boost the effects of heroin and other opioids has made them a favored drug type among the drug-using population [56]. They are also widely used by the COC-using population, especially clonazepam because of its ability to

increase the seizure threshold [56]. They are also used in drug-facilitated crimes [56].

Benzodiazepines may be divided into four categories based on their elimination $t_{1/2}$: (1) ultrashort-acting agents; (2) short-acting agents, with $t_{1/2}$ of <6 hours; (3) intermediate-acting agents, with $t_{1/2}$ of 6–24 hours; and (4) long-acting agents, with $t_{1/2}$ of >24 hours [54] These pharmacokinetic properties, in part, determine the primary clinical applications for some benzodiazepines. For instance, midazolam ($t_{1/2}$, 1–4 hours) is used for preanesthetic sedation or for sedation for endoscopic procedures because of its rapid onset and short duration of action. Benzodiazepines useful in treating anxiety generally have intermediate to long elimination $t_{1/2}$ (alprazolam and diazepam), and those primarily used as anticonvulsants (clonazepam) have the longest. Elimination $t_{1/2}$ is not the sole determinant of duration of action of benzodiazepines, and, in some cases, the rate of drug redistribution from the CNS may be a more important factor [7]. Benzodiazepines undergo hepatic oxidation (phase I) and conjugation (phase II), often forming metabolites with pharmacologic activity (Table 52.5). Following these reactions, conjugation with glucuronic acid occurs; these glucuronidated metabolites constitute the major urinary products of benzodiazepines (Fig. 52.2) [16].

Virtually, all results of the pharmacologic effects of benzodiazepines are caused by their actions on the CNS. The most prominent of these effects are (1) sedation; (2) hypnosis; (3) decreased anxiety; (4) muscle relaxation; (5) anterograde amnesia; and (6) anticonvulsant activity. Only two effects of these drugs result from peripheral actions: (1) coronary vasodilation, which is observed after

TABLE 52.5 Half-life and primary metabolite of select benzodiazepines.

Drug	Half-life (h)	Significant phase I metabolites
Short-acting		
Midazolam	1–4	α-Hydroxy-midazolam
Estazolam	10–24	3-Hydroxy-estazolam
Flurazepam	1–3 47–100 (*N*-desalkyl-flurazepam)	Hydroxy-ethyl-flurazepam *N*-desalkyl-flurazepam[a]
Temazepam	3–13	Oxazepam
Triazolam	1.8–3.9	α-Hydroxy-triazolam
Intermediate-acting		
Flunitrazepam[b]	9–25	7-Amino-flunitrazepam
Long-acting		
Diazepam	21–37	Nordiazepam[a] Oxazepam[a] Temazepam[a]
Quazepam	39–53	3-Hydroxy-quazepam *N*-desalkyl-2-oxo-quazepam 2-Oxo-3-hydroxy-quazepam
Alprazolam	6–27	α-Hydroxy-alprazolam
Chlordiazepoxide	6–27	Nordiazepam[a] Oxazepam[a]
Clonazepam	19–60	7-Amino-clonazepam
Clorazepate[c]	2 31–97 (nordiazepam)	Nordiazepam[a] Oxazepam[a]
Lorazepam	9–16	
Oxazepam	4–11	

[a] *Active metabolite.*
[b] *Not available in the United States.*
[c] *Converted to nordiazepam by gastric HCl.*

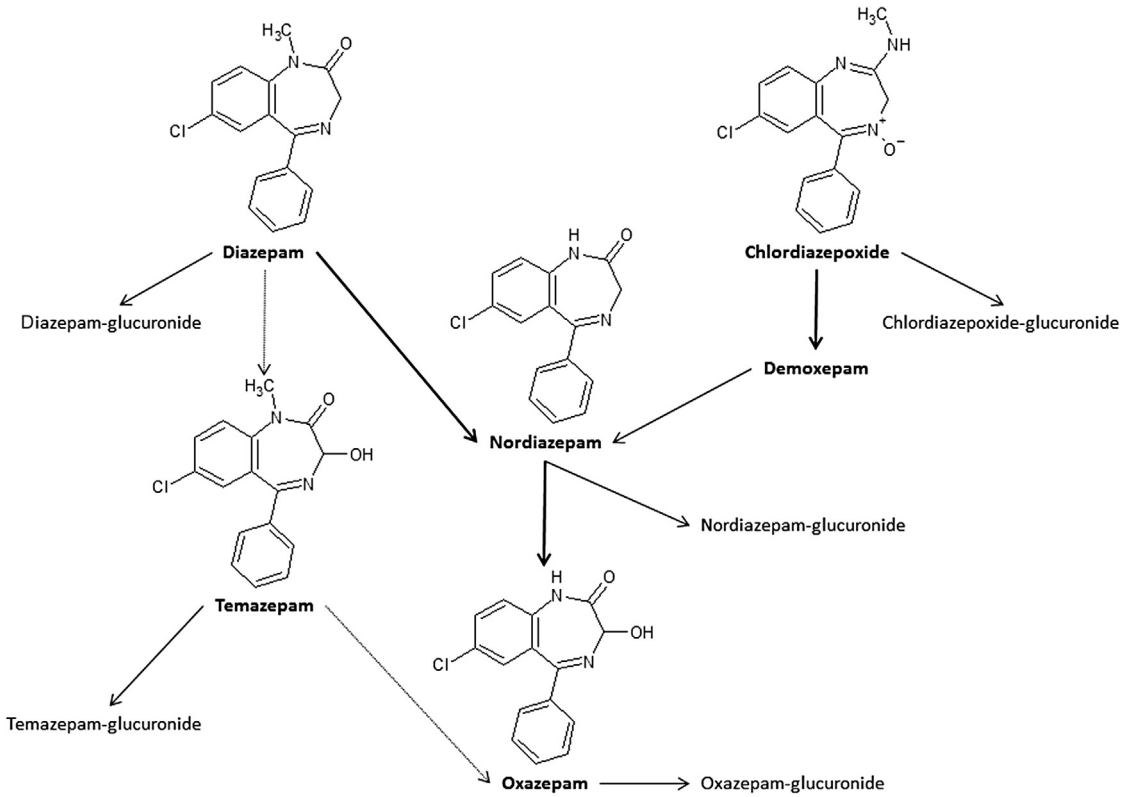


FIGURE 52.2 Metabolic pathway of select benzodiazepines.

intravenous administration of therapeutic doses of certain benzodiazepines; and (2) neuromuscular blockade, observed only after the administration of very high doses. Ethanol increases both the rate of absorption of benzodiazepines and associated CNS depression. Except for additive effects with other sedative or hypnotic drugs, reports of clinically important pharmacodynamic interactions between benzodiazepines and other drugs have been infrequent.

Long-acting benzodiazepines (diazepam, chlordiazepoxide, and clorazepate) are given in relatively large doses and may be detected for several days to weeks or even months following long-term use. Short-acting benzodiazepines (alprazolam and triazolam) are used in lower doses and might be detected only for a few days. The treatment of benzodiazepine toxicity is primarily supportive. Flumazenil may be used in select cases and is a competitive inhibitor of the benzodiazepine site on the GABA complex. It finds its greatest utility in the reversal of benzodiazepine-induced sedation from minor surgical procedures. However, flumazenil should not be administered as a nonspecific coma-reversal drug and should be used with extreme caution after intentional benzodiazepine overdose, because it has the potential to precipitate withdrawal in benzodiazepine-dependent individuals and/or to induce seizures in those at risk.

Benzodiazepines are measured using a variety of techniques. However, their structural diversity and wide variations in potency provide a challenge for laboratories to detect all relevant members in one analytical scheme. Currently, there are several commercial immunoassay systems available for the detection of benzodiazepines and metabolites, but they differ somewhat in their ability to detect each of the respective benzodiazepines, their metabolites, and glucuronide conjugates. Cross-reactivity in screening immunoassays of the various benzodiazepines and their metabolites varies considerably from manufacturer to manufacturer, and screening assays are not able to distinguish between the individual benzodiazepines. Most assays are calibrated to the common metabolites oxazepam, temazepam, or nordiazepam [57]. However, the large number of different functional groups that may be present on the benzodiazepine nucleus makes it difficult to detect all drugs in this class, and some compounds such as lorazepam, clonazepam, midazolam, and flunitrazepam may not be detected by many assays.

GC–MS with derivatization confirmation method is still common [56]; however, LC–MS and LC–MS/MS are becoming increasingly useful and popular because of the ability to detect benzodiazepines and metabolites without derivatization [16]. Typically, benzodiazepines and their metabolites have been extracted from biological

specimens by liquid−liquid extraction or solid-phase extraction. When urine specimens are analyzed, a hydrolysis step to cleave the glucuronide conjugates is frequently used [55]. Enzymatic hydrolysis is preferred over acid hydrolysis, because some benzodiazepines are unstable and rearrange to form benzophenones [55].

Cannabinoids

Cannabinoids are the most commonly used illicit substance in the world and come from the *Cannabis sativa* and *Cannabis indica* plants. Cannabinoids have been used as a medicinal and an illicit psychotropic agent for centuries. Recently, within the United States, several states have independently decriminalized or legalized the use and sale of medical and/or recreational cannabinoids. Delta-9-tetrahydrocannabinol (THC), the primary psychoactive component, is typically consumed by smoking the plant leaves, flower buds, and sometimes stems. THC has also been extracted from the glandular hairs of cannabis flowers and produced as a resin (hashish). Hashish is often a more potent form, and has been mixed into foods, brewed as tea then ingested, or smoked. Hemp oil has also been extracted from cannabis seeds for use in soaps, body care products, and dietary supplements, and is used because of its high essential fatty acid content, but negligible THC content.

The main psychotropic effects of THC are (1) euphoria; (2) distorted perceptions; (3) relaxation; and (4) a feeling of well-being. Additional effects may include feelings of euphoria and relaxation, altered time perception, lack of concentration, impaired learning and memory, and mood changes such as paranoia, psychosis, and panic attacks [58]. Although marijuana is the most frequently used illicit drug, it does have some limited legitimate medicinal use to treat anorexia and nausea in patients with acquired immunodeficiency syndrome and those with nausea and vomiting associated with chemotherapy, or asthma and glaucoma [59]. Recently, the Food and Drug Administration (FDA) has approved cannabidiol (CBD; Epidiolex) to treat seizures associated with Lennox−Gastaut syndrome or Dravet syndrome in patients 2 years of age and older (https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm611046.htm, accessed November 15, 2018).

The main psychotropic effects after inhalation of marijuana occur within minutes and persist for several hours. The peak plasma concentration of THC is dependent on the dose and numerous factors contribute to the variability in dose, such as (1) method of consumption; (2) depth of inhalation; (3) exposure frequency; and (4) cannabis potency [58]. Onset of clinical symptoms and peak plasma concentrations after oral ingestion of THC is slower than after inhalation, primarily as the result of first-pass hepatic clearance, and two THC peaks are frequently observed because of enterohepatic circulation [58]. THC is metabolized by CYP450 liver enzymes to >100 metabolites. The main active metabolite, 11-hydroxy-delta-9-THC, is further oxidized to the most abundant inactive 11-nor-9-carboxy-THC (THC-COOH) [58] Some researchers have used the concentration and ratio of THC and its metabolites to predict acute versus chronic exposure to marijuana [60]. Epidiolex is administered and an oral solution of 100 mg/mL of CBD [61]. It has a T_{max} of 2.5−5 hours, is highly protein bound, and has an elimination $t_{1/2}$ of 56−61 hours [61]. CBD is metabolized primarily in the liver by CYP2C19 and CYP3A4 enzymes, into an active metabolite 7-OH-CBD, which is further oxidized to an inactive metabolite 7-COOH-CBD molecule [61].

In light of recent and ongoing legislation regarding medical and recreational cannabis, interest in monitoring cannabis use and abuse has increased. An immunoassay method is typically used to screen for potential cannabinoid use. Immunoassay screens have been designed to detect cannabis use in urine samples using antibody reagents developed against the inactive THC-COOH metabolite; these reagents cross-react with numerous other THC metabolites. A presumptive positive sample should be confirmed by quantitative GC−MS or LC−MS/MS. Concerns have been raised with regards to the potential for false-positive results from dietary sources and "passive inhalation" of sufficient side-stream marijuana smoke from nearby users. Numerous studies have been conducted to investigate exposure to THC from second-hand smoke, concluding that the SAMHSA cutoff (THC-COOH 50 ng/mL cutoff, immunoassay) is thought to be sufficient to separate moderate passive exposure from first-hand inhalation exposure to THC [62,63]. Therefore it is improbable that a passive inhaler would be able to sustain exposure to significant THC concentrations long enough to produce a positive drug screen.

A positive result from a urine cannabinoid screen or confirmation does not indicate intoxication or degree of exposure. The window of detection for the urine concentration of THC-COOH varies among casual (2−7 days) and chronic abusers (up to 73 days) of marijuana and is dose dependent [7]. Variables affecting the duration of detection include (1) dose; (2) frequency of exposure; (3) route of exposure; (4) body composition; (5) fluid excretion; and (6) method of detection. Therefore monitoring of abstinence is particularly challenging. Dilution of urine due to normal biological fluctuations (hydration) or ingested adulterants can cause a negative result one day and a positive on the next. To correct for hydration fluctuations, urine concentrations of THC-COOH per milligram creatinine are normalized for monitoring individuals who are resuming cannabis use. Some people

suggest that, using these normalized THC-COOH:creatinine concentrations, a ratio is calculated by comparing any normalized urine specimen (U2) with a previously collected normalized urine specimen (U1). "New use" is defined as a U2/U1 ratio of $\geq 0.5-1.5$ collected from urine specimens taken more than 24 hours apart and containing THC-COOH concentrations of >15 ng/mL [64–66]. Using the 1.5 cutoff rate results in decreased false-positives, but increased false-negative decisions [64]. Another model has been proposed to distinguish new versus residual use in chronic daily users [67].

Monitoring minor cannabinoids present in cannabis and their metabolites has also been proposed as a mechanism to distinguish recent use from prolonged elimination. Several minor cannabinoids, such as CBD, cannabinol, cannabigerol, tetrahydrocannabivarin (THCV), and its metabolite, 11-nor-9-carboxy-THCV, have been proposed as possible markers of recent cannabis intake in different biological matrices [68–73]. Finally, because of the use of CBD in epilepsy, numerous studies are evaluating the potential for a more traditional therapeutic drug management style approach to monitoring of CBD concentration in blood or serum of patients taking this drug to control seizures.

Synthetic cannabinoids

Synthetic cannabinoids are not naturally occurring compounds, but synthetic compounds originally developed as potential therapeutic compounds used to characterize cannabinoid receptors (CB-1 and CB-2). Numerous compounds have been described in the scientific literature, and clandestine labs continue to substitute functional moieties and synthesize new compounds. Synthetic cannabinoids are commonly sprayed on dried plant material and smoked. The clinical effects of these agents have not been well defined, with most clinical reports based on patient history of what was taken versus what was truly present in the product abused. Clinical effects of synthetic cannabinoids may be similar to THC. Yet some reported effects, such as anxiety, agitation, psychosis, and suicidality, may be more prominent.

While there are commercial enzyme-linked immunoassays available, no single immunoassay kit can detect all known synthetic cannabinoid compounds. GC–MS, LC–MS/MS, LC–HRMS, or TOF may be utilized to screen and/or confirm the growing release of new synthetic cannabinoids entering the market, but labs have to consistently update the targets, as new compounds appear quickly in an attempt to outpace detection and legislation. Furthermore, the metabolism of synthetic cannabinoids is complex, such that many of the compounds have common metabolites, and confirmatory testing should not only be able to detect the large number of metabolites, but also be adaptable to accommodate the rapidly changing landscape of drugs being introduced.

Cocaine

COC is an alkaloid found in *Erythroxylon coca*, which grows principally in the northern South American Andes and to a lesser extent in India, Africa, and Java [74]. In clinical medicine, it has been used mainly for local anesthesia and vasoconstriction in nasal surgery, and to dilate pupils in ophthalmology. COC is still one of the most common illicit drugs of abuse, and, according to the National Survey on Drug Use and Health, the 2017 rate of use for COC (powder and crack combined) among individuals aged 12 and older has remained relatively stable since 2002 (https://www.samhsa.gov/data/nsduh/reports-detailed-tables-2017-NSDUH, accessed June 28, 2019).

COC is sold on the street in two forms: a hydrochloride salt (powder) and a free-base product known as "crack." The hydrochloride salt form of COC is administered by nasal insufflation ("snorting") or intravenously. "Crack" is a free-base form that has not been neutralized by an acid to make the hydrochloride salt. It comes as a rock crystal that is heated and its vapors smoked. The term refers to the crackling sound heard when it is heated [74].

The metabolism of COC is complex and occurs via both nonenzymatic hydrolysis and enzymatic transformation in the plasma and liver, where it is rapidly metabolized to benzoylecgonine (BE) and ecogonine methyl ester, both of which are inactive [74]. COC contains two ester moieties; the alkyl ester is hydrolyzed to its major metabolite BE via spontaneous hydrolysis at physiologic and alkaline pH [75]. It has been shown that COC is also hydrolyzed to BE by liver carboxylesterases [76]. BE is considered to be a pharmacologically inactive metabolite, but because its $t_{1/2}$ is longer than that of COC, it is the most commonly monitored analyte in urine for determination of COC use. BE is further metabolized to minor metabolites such as *m*-hydroxybenzoylecgonine (m-HOBE) and *p*-hydroxybenzoylecgonine [74]. Of these, m-HOBE has been shown to be an important metabolite in the meconium of COC-exposed babies [77,78]. Norcocaine (NC) is an *N*-demethylated metabolite of COC produced by liver cytochrome P450; it is of clinical interest because of its conversion into hepatotoxic metabolites [74]. Anhydroecgonine methyl ester (AEME; methyl ecgonidine) has been identified as a unique COC metabolite after smoked COC ("crack") administration. Anhydroecgonine ethyl ester (AEEE; ethyl ecgonidine) has been identified in COC smokers who also use ethyl alcohol [79].

COC has cardiovascular effects and is a potent CNS stimulant that elicits a state of increased alertness and euphoria with pharmacological actions similar to those of amphetamine but of shorter duration. These CNS effects

are thought to be largely associated with the ability of COC to block dopamine reuptake at nerve synapses, thereby prolonging the action of dopamine in the CNS [74]. COC also blocks the reuptake of norepinephrine at presynaptic nerve terminals; this produces a sympathomimetic response that may result in (1) mydriasis; (2) diaphoresis; (3) hyperactive bowel sounds; (4) tachycardia; (5) hypertension; (6) hyperthermia; (7) hyperactivity; (8) agitation; (9) seizures; or (10) coma. Sudden death due to cardiotoxicity may occur following COC use. Death may also occur following the sequential development of hyperthermia, and respiratory arrest. Excited delirium and extreme physical activity may lead to rhabdomyolysis, acute renal failure, and disseminated intravascular coagulopathy [74].

COC is effective as a local anesthetic and vasoconstrictor of mucous membranes and, therefore, is used clinically for nasal surgery, rhinoplasty, and emergency nasotracheal intubation. COC is also frequently used with other drugs recreationally, most commonly ethanol. In simultaneous COC and ethanol use, liver methylesterase catalyzes the conversion of COC to BE and the transesterification of COC to cocaethylene (CE) [74].

The elimination $t_{1/2}$ for COC ranges from 0.5 to 1.5 hours, and for BE from 4 to 7 hours [7]. Only small amounts of COC are excreted in urine. The elimination $t_{1/2}$ for CE is 2.5−6 hours [7], which is considerably longer than that for COC. BE excretion is detectable for 1−3 days following COC use. However, for chronic heavy COC users, the detection time may extend to 10−22 days following the last dose [16], apparently because of tissue storage of COC.

Due to the above mentioned longer elimination $t_{1/2}$ of BE, it is the analyte of choice in screening for COC use [74]. The initial screening test for BE is typically immunoassay, and screening immunoassays frequently have a 300-ng/mL or 150-ng/mL cutoff for it. Most confirmation assays offer quantification of both the parent drug and metabolite. Numerous methods have been described for the measurement of COC and its metabolites. GC techniques for analysis of COC and its metabolites require derivatization, and GC−MS methods are still frequently the method of choice [16]. LC−MS/MS methods have also been described, including COC, BE, and m-HOBE, along with other relevant secondary metabolites, such as CE, NC, AEME, and AEEE [80].

Lysergic acid diethylamide

Lysergic acid diethylamide (LSD) is synthesized from D-lysergic acid, a naturally occurring ergot alkaloid found in the fungus *Claviceps purpurea*, which grows on wheat and other grains. LSD shares structural features with serotonin (5-hydroxytryptamine; a major CNS neurotransmitter and neuromodulator) [81]. LSD binds to serotonin receptors in the CNS and acts as a serotonin agonist. The principal psychological effects of LSD are perceptual distortions of color, sound, distance, and shape (synesthesia); depersonalization and loss of body image; and rapidly changing emotions from ecstasy to depression or paranoia. The physiologic effects of LSD are related to its sympathomimetic actions and include mydriasis (most frequent and consistent), tachycardia, increased body temperature, diaphoresis, and hypertension; at higher doses, parasympathomimetic actions may be observed [e.g., salivation, lacrimation, nausea, and vomiting (muscarinic actions)]. Neuromuscular effects may include paresthesia, muscle twitches, and incoordination (nicotinic actions) [81].

LSD is used illicitly because of its hallucinogenic effects. The most common adverse effects of LSD are panic attacks. In addition, unpredictable recurrence of hallucinations (flashbacks) may occur weeks or months after last drug use, and LSD may elicit psychotic reactions. No evidence suggests that repeated LSD use results in dependence or withdrawal symptoms [81].

Popular dosage forms include powder, gelatin capsule, tablet, and LSD-impregnated sugar cubes, filter paper, or postage stamps. The drug is rapidly absorbed from the GI tract; the effects begin within 40−60 minutes, peaking at ∼2−4 hours, and subside by 6−8 hours. The elimination $t_{1/2}$ is ∼3 hours. The metabolism of LSD in humans is incompletely understood, but 2-oxo-3-hydroxy-LSD is present in urine at concentrations of 10- to 43-fold greater than LSD [82]. *N*-demethyl-LSD is also present in urine specimens, but at concentrations approximately equivalent to those of LSD [83].

The typical dose is low (20−80 μg) and is rapidly metabolized, and only ∼1%−2% of the drug is excreted unchanged in urine. Thus detection of LSD presents an especially difficult analytical challenge. Even with sensitive assays, the detection window for LSD is generally only 12−24 hours [84]. Immunoassays are targeted to detect LSD, and confirmation assays may be performed by GC−MS, GC−MS/MS, LC−MS/MS, or LC−MS [16].

Opioids

The term *opioid* describes a wide range of compounds encompassing the natural and semisynthetic opiates—essentially variations on the structure of morphine—and fully synthetic opioids with minimal structural homology to the natural alkaloids [85]. The defining characteristic of this class of drugs is their interaction with opioid receptors [86].

For pain management, opioids are used frequently in treating acute needs, such as postsurgical analgesia, and in relieving moderate-to-severe chronic pain (see later section). However, most opioids are highly addictive and

have potentially life-threatening side effects; thus the benefits of their use must be carefully weighed against the chance of serious consequences. In addition, the development of tolerance and the risk of prescription diversion complicate even further the process of monitoring long-term opioid therapy for compliance and efficacy.

The hallmark of opioids is their ability to interact with the family of opioid receptors that are variably distributed throughout the body. These include mu-opioid receptor, delta-opioid receptor, and kappa-opioid receptor (MOR, DOR, and KOR, or μ, δ, and κ, respectively). Opioid receptor agonists typically produce analgesia, and antagonists block this response [86]. The metabolism of opioids is varied, but numerous biotransformations are common to these drugs (Table 52.6). Several of the most commonly used opiates are formed in vivo by metabolism of other compounds, as is seen with codeine demethylation, resulting in conversion to morphine [86]. This interconversion is a frequent source of confusion and must be considered when the results of opiate screens are interpreted (Fig. 52.3).

Natural opium alkaloids

Morphine

The archetypical opiate, morphine, is used as the basis of comparison for relative characterizations of the opioid class. Morphine interacts primarily with MOR to mediate its effects, but it also shows some affinity for KOR [86]. Its major metabolites are glucuronide conjugates, including inactive morphine-3-glucuronide ($\sim$60% of metabolite), active morphine-6-glucuronide (M6G; $\sim$10% of metabolite), and a small amount of morphine-3,6-diglucuronide [86].

With long-term administration and when morphine concentrations are high, a minor fraction is converted to hydromorphone (up to 2.5% of the urine morphine concentration) [16]. M6G has greater MOR agonist activity than morphine and appears to contribute less to unwanted side effects [86]. The elimination $t_{1/2}$ for glucuronides is longer than for morphine. Therefore glucuronides accumulate in serum to greater concentrations than morphine, and, in patients with renal insufficiency, morphine glucuronides are thought to significantly contribute to opioid toxicity [86].

Codeine

Codeine shows poor affinity for MOR and has only about one-tenth the analgesic potency of morphine [16]. Analgesia is attributed to the small fraction ($<$10%) of codeine converted to morphine by CYP2D6 via *O*-demethylation; therefore it is generally considered a prodrug. Both codeine and morphine may be detected in urine following codeine ingestion; however, after 30 hours, only morphine may be detectable [16]. Long-term, high-dose administration leads to metabolism to hydrocodone (up to 11% of the urine codeine concentration) [87].

Genetic variation may play a significant role in the metabolism of codeine and several other opioids. More than 130 alleles have been described for CYP2D6, with resultant enzymatic activity ranging from essentially zero, in the case of null alleles, to many times higher than normal, in the case of amplified alleles (https://www.pharmvar.org/gene/CYP2D6, accessed June 28, 2019). Thus, at the same codeine dose, patients with minimal CYP2D6 activity (poor metabolizers) would likely receive inadequate analgesia because of lack of conversion to morphine; however, patients with very high CYP2D6 activity

TABLE 52.6 Half-life and primary metabolite of select opioids.

Drug	Half-life [16]	Significant phase I metabolites
Natural opium alkaloids		
Morphine	1.3–6.7 h	
Codeine	1.2–3.9 h	Morphine
Semisynthetic opiates		
Heroin	2–6 min	6-Monoacetymorphine morphine
Hydrocodone	3.4–8.8 h	Hydromorphone
Hydromorphone	3–9 h	
Oxycodone	3–6 h	Oxymorphone
Oxymorphone	4–12 h	

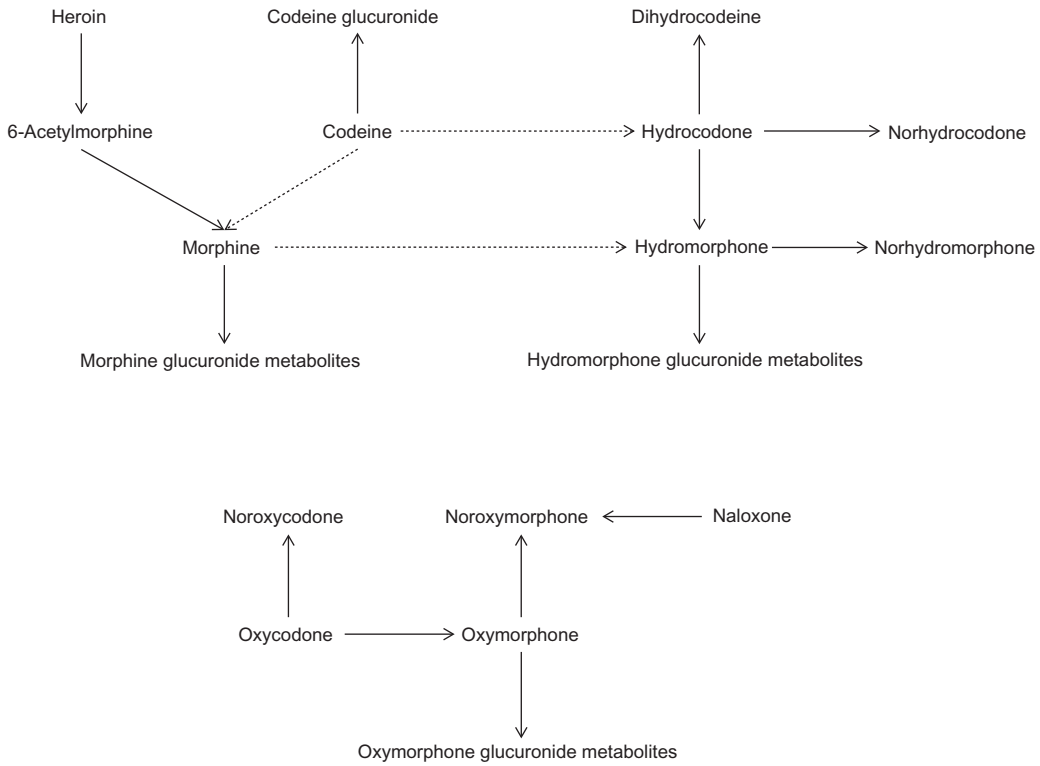


FIGURE 52.3 Metabolic pathway of opiates.

(ultrarapid metabolizers) would be at risk for adverse responses to excessive morphine [86].

Semisynthetic opiates

Heroin

Heroin is a synthetic opiate that is made from morphine and is also called *diacetylmorphine*; it has an analgesic potency two-to-three times that of morphine, possibly because of increased penetration across the blood−brain barrier. Heroin itself is rarely found in body fluids because of its extremely short $t_{1/2}$ (2−6 minutes); the metabolite, 6-acetylmorphine, is hydrolyzed to morphine; although it has a longer $t_{1/2}$ (6−25 minutes), it is detectable in urine only for ∼8 hours after administration [16]. 6-Monoacetylmorphine (6-MAM) is also four-to-six times more potent than morphine. Other than the presence of heroin's unique metabolite 6-MAM, the metabolic profile of heroin resembles that of morphine. Given that acetylcodeine is a common contaminant of heroin, both morphine and low concentrations of codeine are frequently detected in urine following heroin use.

Hydrocodone and hydromorphone

Hydrocodone has approximately six times the potency and greater oral bioavailability than codeine [16]. Hydrocodone is *O*-demethylated to hydromorphone, and *N*-demethylated to form norhydrocodone. Similar to codeine, hydrocodone is metabolized by CYP2D6 to an active metabolite (hydromorphone) and therefore may be subject to pharmacogenetic variability in patients with abnormal CYP2D6 activity [88]. The clinical impact of this is unclear. It has been suggested that most of the pharmacologic effects of hydrocodone actually result from the hydromorphone formed during metabolism. However, studies are somewhat contradictory [16]. On the other hand, hydromorphone is five-to-seven times more potent than morphine and can be used as an analgesic by itself. Similar to morphine, hydromorphone is metabolized in large part to a 3-glucuronide [16].

Oxycodone and oxymorphone

Oxycodone is a potent analgesic with high oral bioavailability [86] that is frequently formulated in combination with acetaminophen. Therefore the detection of acetaminophen along with oxycodone in the urine of patients who display an opiate toxidrome should lead to measurement of serum acetaminophen to assess toxicity. Although its own strong analgesic activity precludes oxycodone from being considered a prodrug, it is converted to a highly active metabolite, oxymorphone, through CYP2D6 activity [16]. This conversion appears to be less of a concern for CYP2D6 poor metabolizers, in whom oxycodone itself still provides analgesia, than for ultrarapid metabolizers,

who could be at increased risk for adverse effects [89]. Oxymorphone also provides potent analgesia. The majority of oxymorphone is metabolized to the 3-glucuronide; a minor metabolite, 6-hydroxyoxymorphone, is also active [16].

Fully synthetic opioids

Fentanyl

Fentanyl has a rapid onset and short duration of effect. Fentanyl provides the structural backbone for a number of related, ultrashort-acting opioids, including remifentanil and sufentanil. Norfentanyl, the primary metabolite, is inactive [90]; the high potency of fentanyl and the clinical insignificance of its metabolites make it a preferred analgesic for patients with major organ failure [91]. Transdermal fentanyl patches are used for long-term administration and are gaining popularity among drug abusers, even though nonstandard application of the patch (e.g., chewing and extraction) carries substantial risk for overdose [92].

Meperidine

Meperidine has analgesic potency comparable with or somewhat lower than that of morphine [93]. One major metabolite, normeperidine, also has analgesic activity; normeperidine is thought to be responsible for the serotonergic toxicity of meperidine, particularly in patients receiving concomitant monoamine oxidase inhibitors [86]. Meperidine use has declined in recent years in favor of alternatives.

Methadone

A relatively long-acting opiate, methadone, is used both for analgesia and in the treatment of opioid addiction [86]. It is thought to provide: (1) milder withdrawal; (2) somewhat lower potential for abuse; and (3) reduced exposure to the risks of illicit intravenous drug use. Methadone has affinity for both MOR and DOR [94], the latter of which may explain its apparent utility in patients whose pain no longer responds to other opioids [86]. Substantial interindividual and intraindividual variability in metabolism and elimination has been noted; both urine pH and seemingly self-inducible metabolism substantially influence the pharmacokinetics of this compound, as do commonly administered drugs such as benzodiazepines and antiretrovirals [86]. Although a large fraction of methadone is excreted unchanged, measurement of a metabolite, such as 2-ethylidene-1,5-dimethyl-3,3-diphenylpyrrolidine (EDDP), in the setting of addiction treatment provides evidence for patient compliance rather than an exogenously spiked sample. EDDP excretion is less pH-dependent than is clearance of the parent drug [16,86]. Use of the methadone/EDDP ratio to assess compliance has been suggested but is complicated by the pharmacokinetic variability already described.

Tramadol

Unlike the majority of opioid agonists, tramadol has low abuse potential and therefore is unscheduled. It has low affinity for opioid receptors and mediates analgesia through opioid-independent regulation of neurotransmitter uptake; however, its main active metabolite (*O*-desmethyltramadol) is a potent opioid receptor agonist [87]. These mechanisms are thought to work synergistically to provide greater total pain relief than the sum of each individual component. Metabolism to *O*-desmethyltramadol occurs via CYP2D6; thus opioid-like effects are subject to genetic variability [95]. However, because of its effects on neurotransmission, tramadol has the potential to cause serotonergic toxicity even in patients lacking CYP2D6 [87].

Opioid antagonists and mixed agonist/antagonists

These clinically useful compounds produce very different physiologic responses, depending on the situation. For example, in opioid-naive patients, mixed agonist/antagonists (MAAs) provide MOR-mediated analgesia with less risk of an adverse reaction, but the same dose in an opioid-tolerant patient may precipitate immediate withdrawal. In medical usage, coadministration of low-dose antagonists or MAAs alleviates minor opioid-induced side effects and appears useful in preventing opioid tolerance. In opioid addiction treatment, the addition of a low-dose antagonist to maintenance therapy seems to minimize subjective "feel-good" effects without substantially worsening withdrawal symptoms.

Buprenorphine

Buprenorphine is an MOR partial agonist and a KOR antagonist. Low doses provide analgesia through MOR activation, but, unlike full agonists, pain relief has a maximal threshold or "ceiling effect" [96]. Buprenorphine is available as sublingual tablets (with or without naloxone) for the treatment of opioid dependence and as a transdermal patch [97]. Buprenorphine is metabolized via *N*-dealkylation by CYP3A4 to the active compound, norbuprenorphine, both of which can be further conjugated to inactive metabolites [98,99].

Naloxone

The prototypical opioid antagonist, naloxone, binds nonspecifically to all three receptor types, with the greatest

effect at MOR and the least effect at DOR [86]. Naloxone is commonly used in comatose patients as a therapeutic and diagnostic agent. When given in appropriate amounts in patients who are comatosed because of opiate intoxication, the clinician will see restoration of respiratory function, and improved level of consciousness. Its clinical efficacy lasts for as little as 45 minutes. Therefore patients are at risk for recurrence of narcotic effect. This is particularly true for patients exposed to opioids with long elimination $t_{1/2}$, such as methadone and sustained-release opioid products. Its efficacy is much greater by intravenous administration as compared with oral and sublingual routes [97]. This characteristic is advantageous in deterring misuse of prescribed opioids: oral or sublingual opioid/naloxone formulations provide the desired benefit when taken properly, but when diverted for intravenous use cause opioid antagonism and may precipitate withdrawal [97].

Opioid analysis

Given their relatively rapid TAT and ability to identify several opiates, immunoassays are commonly used to screen for opiates. For clinical application, a cutoff of 300 ng/mL morphine (or morphine equivalents) is commonly used to distinguish negative from positive urine specimens, whereas a cutoff of 2000 ng/mL is common for workplace drug screening. Antibodies in opiate screens commonly target morphine, because commercial immunoassay development has largely been driven by detection of illicit heroin use. Wide variability in cross-reactivity to other opiates has been noted; thus some opiates or opioids with high abuse potential, such as oxycodone, are often poorly detected [16]. To address this problem, several immunoassays are commercially available for individual opioids, such as fentanyl, methadone, and buprenorphine, which do not cross-react with the opiate immunoassay at all.

Phencyclidine and ketamine

On the street, PCP and ketamine are sold under a variety of names. They are available as a colorless, odorless liquid, or as a white powder. PCP is listed as a Schedule II drug in the US Federal Controlled Substance Act and is not approved for human use. Ketamine is a Schedule III drug, used as an anesthetic for short surgical procedures. Both drugs have been used illicitly in human cases of drug abuse, as well as in cases of drug-facilitated crimes. Ketamine and PCP share similar structural features and pharmacologic actions [83]. They are classified as dissociative anesthetics, because they produce rapid-acting dissociation of perception, consciousness, movement, and memory [83,100]. The effects are dose dependent and vary between individuals. Some individuals experience effects similar to the psychosis observed in schizophrenia. An anesthetic dose produces profound analgesia, where the individual is awake yet incapacitated, with limited voluntary limb movement. Both PCP and ketamine have also been associated with psychological disturbances.

PCP [1-(1-**p**henyl**c**yclohexyl)-**p**iperidine] is used recreationally for its mind-altering "out of body" experience. Recreational use of PCP declined in the 1980s but has reemerged in recent years. Presentation of adverse effects, such as dysphoria, ataxia, nystagmus, agitation, anxiety, paranoia, amnesia, seizures, muscle rigidity, hostility, delirium, delusions, and hallucinations, is unpredictable. Evidence of flashbacks has been documented with PCP use [100]. With repeated use of PCP, psychological dependence may develop, but tolerance or withdrawal syndrome is not profound. A sense of superhuman strength coupled with lack of pain perception may lead to excessive physical exertion and accidental or intentionally induced trauma. Thus PCP-related deaths most often are secondary to these adverse behavioral drug effects. Treatment of PCP toxicity is supportive [100].

Ketamine was discovered during subsequent studies characterizing PCP analogs. The liquid or powder form can be easily disguised in a victim's beverage; this has resulted in its use in drug-facilitated crimes. Ketamine powder can even be sprinkled onto marijuana or tobacco and smoked. It produces effects similar to those of PCP. Ketamine has a $t_{1/2}$ of $\sim 2-3$ hours [16] and is metabolized to norketamine, which has about one-third the activity of ketamine, and to dehydronorketamine, which may also be active [83]. Duration of anesthetic effects is dose dependent (usually <1 hours), and effects on the senses, judgment, and coordination can have a longer duration ($\approx 6-24$ hours). At higher doses, ketamine causes delirium, amnesia, dissociative anesthesia, hallucinations, delirium, hypersalivation, nystagmus, impaired motor function, hypertension, and potentially fatal respiratory problems. Effects on blood pressure and respiratory depression are significantly enhanced when coingested with alcohol.

Initial screening is typically done by immunoassay. Whether or not PCP and ketamine are included in a general urine drug screen depends on applicable regulations and on the prevalence of use in the local community. In some locations, the prevalence of PCP or ketamine use may be too low to warrant routine screening. Immunoassays are generally reliable; however, for PCP, false-positives have been reported because of high concentrations of dextromethorphan, diphenhydramine, and thioridazine [16]. Immunoassay-positive specimens should be confirmed, most commonly using GC−MS or LC−MS/MS analyses [16].

Specimen validity testing

Specimen validity testing is primarily performed to ensure a valid urine specimen was collected for drug testing. Several techniques can be used by people attempting to mask or adulterate drug use to avoid detection. These tactics may include the exchange of urine from a drug-free individual; or dilution of the urine specimen by excessive consumption of water, use of a diuretic or simple addition of water to the specimen to reduce drug concentrations to below cutoff limits. Also, readily available substances, such as detergent, bleach, salt, alkali, ammonia, tetrahydrozoline, or acid, may be added to the specimen after collection in an attempt to interfere with immunoassay screening procedures. Other more sophisticated adulterants specifically marketed to avoid drug detection include glutaraldehyde (Urine Aid; Clear Choice), nitrite (Klear; Whizzies), chromate (Urine Luck; Sweet Pee's Spoiler), and a combination of peroxide and peroxidase (Stealth). These adulterants also interfere with immunoassays to variable degrees, and the oxidizing agents (nitrite, chromate, and peroxide/peroxidase) may result in destruction of morphine, codeine, and the principal metabolite resulting from marijuana use, thus interfering with GC−MS confirmation and immunoassays [101].

Direct observation of urine collection is the most stringent means to guard against specimen exchange or adulteration. However, an individual's right to privacy and dignity must be weighed against the need for the highest degree of certainty of specimen integrity. Alternative measures to prevent specimen adulteration include (1) limitations on clothing or other personal belongings allowed in the specimen collection area; (2) addition of coloring agent to toilet water; and (3) inactivation of the hot water tap. In addition, several validity checks for specimen integrity may be made at the collection site and at the testing site. Validity testing criteria have been established by the Department of Health and Human Service for the drug testing program mandated for US federal employees (https://www.transportation.gov/odapc/part40, accessed July 2, 2019).

Pain management

UDT is an essential tool and recommended by numerous clinical practice guidelines for pain management. In the United States, in the late 1990s, awareness grew about the personal and economic costs of poorly controlled pain. Billions of dollars were being spent each year to treat between 30- and 80-million Americans with debilitating, chronic pain [102]. Congress reacted by passing into law the "Decade of Pain Control and Research." Overnight, pain became the fifth vital sign and the successful treatment of pain synonymous with quality of care.

In short order, the negative consequences of widespread opioid use surfaced. Since 1999, the number of opioid-related deaths in the United States has more than doubled and now exceeds the number of deaths related to heroin and COC abuse combined [103−105]. In 2012, drug overdose was the leading cause of injury death, surpassing that from motor vehicle crashes between the ages 25 and 64 [106]. For every one death from opioid abuse, there are 35 emergency department visits, 161 cases of abuse, and 461 cases of misuse, contributing to more than $55 billion dollars in society and work-related costs [107].

UDT in clinical pain management practice is fundamentally different than forensic drug testing. Forensic testing follows chain of custody, and the expectation is that the results for illicit and controlled substances will be negative. In virtually, all clinic pain medicine practices, no chain of custody collection procedures are used, and the controlled medication that patients are prescribed should be present if patients are compliant. The absence of prescribed medications is as much of a professional and ethical conundrum as the presence of illicit or unexpected controlled substances. In addition, the pain physician must be keenly aware that a urine sample yielding consistent results ensures only that the expected parent compound and metabolites are present; not that the patient is using the medication as prescribed. This point cannot be overlooked, as medication nonadherence in the chronic pain population is between 30% and 50% [108].

UDT includes screening tools, such as point of care (POC) and laboratory-based immunoassay testing, as well as confirmatory testing. POC testing is advantageous, because results can be obtained during the clinic visit and could be used for high-risk patients who have to travel far. POC testing could be used at the initial visit where inconsistent results or finding other illicit substances could influence the decision to initiate opioid therapy, or not [109]. However, changes in Medicare reimbursement for POC testing has made it less attractive in the private setting, and its use will likely decrease [110]. Instead, more comprehensive, laboratory-based definitive tests will be recommended for routine compliance monitoring for pain management patients.

Overall, laboratory-based and POC-based immunoassays have moderate sensitivity and low specificity [111]. They provide information about the general drug classes present in the urine but cannot distinguish between drugs within the same class. Immunoassays may completely miss semisynthetic or synthetic opioids, and therefore should be carefully interpreted based on the known limitations of the specific immunoassay used. Therefore confirmatory testing with GC/LC should be considered when screening results are inconsistent with prescribed medications and incorporated as a random part of routine monitoring. The higher cutoff values inherent to most

immunoassays also result in high false-negative rates and may cause the physician to miss aberrant use of prescriptions medication or falsely accusing a patient of diversion. However, in some settings, confirmatory testing can cost in excess of $100 per drug class; therefore restraint should be used.

Recently, numerous algorithmic approaches to UDT in pain medicine have been suggested. While finer details differ, there are a few clear messages [105,107,109,111]:

1. UDT should be used to monitor all patients on chronic opioid therapy.
2. The patient should be informed of and willing to participate in random UDT as documented at the initiation of therapy, in the opioid prescribing contract.
3. Random UDT testing is preferred to scheduled testing, as patients can and will modify usage if told in advance.
4. UDT prior to the initial consultation or POC testing during the initial visit might aid in the decision to initiate opioid therapy.
5. It is important to stratify patient risk into low and high or low, medium, and high categories based on initial screening, comorbid conditions, previous drug/alcohol abuse, and family history.
6. At a minimum, UDT should be performed once per year.
7. More frequent UDT is recommended for patients who are medium-to-high risk.
8. All unexplained results based on urine drug screening should undergo confirmatory testing.
9. Confirmatory testing should be performed at least once per year.
10. Urine temperature, pH, creatinine content, and specific gravity should also be tested to help identify adulterated samples.
11. *Inconsistent* or unexplained results should be discussed with the performing lab and then with the patient in an open, truthful manner.
12. In the event that misuse, abuse, or diversion is suspected, the physician must choose to continue the current treatment plan, take corrective action in an attempt to retain the patient, or dismiss the patient from practice.
13. Corrective actions include counseling, decreasing the time interval between appointments, limiting the number of pills, and referring to an addiction specialist.
14. Dismissing a patient from practice should be based on a prewritten policy that is reviewed with the patient.

In January 2018, the executive summary of the American Association of Clinical Chemistry Laboratory Medicine Practice Guideline (LMPG)—Using Clinical Laboratory Tests to Monitor Drug Therapy in Pain Management Patients was published [112]. The purpose of this LMPG was to provide clinical and laboratory practice recommendations developed using evidence-based approaches to address specific questions regarding the appropriate use of diagnostic tests to monitor adherence or compliance to controlled substances in the pain management population. Ultimately, this guideline produced 26 evidence-based and seven consensus-based recommendations on the use of laboratory and POC tests to monitor the use of controlled substances in pain management patients. One of key evidence-based recommendations is that first-line definitive testing (either qualitative or quantitative) is recommended for detecting the use of relevant over-the-counter medications, prescribed and nonprescribed drugs, and illicit substances in pain management patients. In addition, specimen validity testing (e.g., pH) is recommended, since it is an effective tool to ensure that laboratory test results are correctly interpreted in pain management patients. Specimen validity testing determines the suitability of the urine specimen collected/received, which directly affects the ability to correctly identify relevant over-the-counter medications, prescribed and nonprescribed drugs, and illicit substances used by pain management patients

In the end, UDT is an important tool in clinical pain medicine. The current epidemic of prescription medication abuse underscores the need for a consistent approach to monitoring that utilizes both subjective and objective measures. While some advocate that routine UDT should include all testable substances with abusive potential, such practices are costly and likely unnecessary in many cases. High-risk patients, classified by using validated clinical screening tools, require more careful and more frequent monitoring. POC and other EIA screening methods cannot differentiate between medications within the same drug class and may completely miss semisynthetic and synthetic opioids. For this reason, confirmatory testing with GC/LC should be performed at least annually and on all inconsistent samples. If misuse, abuse, or diversion is suspected, the physician must act in a way that promotes patient health and safety, while abiding by state and federal law. Such actions should be openly discussed and explicitly written in the opioid prescribing contract, signed at the initiation of every opioid treatment plan.

Drugs of abuse testing using other types of specimens

The collection of biological samples for the purpose of determining exposure to various agents is dominated by blood and urine. Blood is considered invasive, and the collection of urine may require some invasion of privacy and loss of dignity; urine specimens are also subject to

adulteration or manipulation to evade detection. For these reasons, alternate biological specimens have been investigated.

Meconium

Illicit drug use during pregnancy is a major social and medical issue. Drug abuse during pregnancy is associated with significant perinatal complications, including a high incidence of (1) stillbirth; (2) meconium-stained fluid; (3) premature rupture of the membranes; (4) maternal hemorrhage (abruptio placentae or placenta previa); and (5) fetal distress [113]. In the neonate, the mortality rate and morbidity (e.g., asphyxia, prematurity, low birth weight, hyaline membrane disease, infection, aspiration pneumonia, cerebral infarction, abnormal heart rate and breathing patterns, and drug withdrawal) are increased [113].

Unfortunately, identification of the drug-exposed mother or her neonate is not easy. Maternal admission of the use of drugs is often inaccurate, principally because of denial about addiction or fear of the consequences stemming from such admission. Likewise, many infants who have been exposed to drugs in utero may appear normal at birth and show no overt manifestations of drug effects. Drug testing is an objective means of determining drug exposure in both mother and infant. Urine testing of the mother or newborn can detect only recent drug use (within a few days before birth), and urine collection from newborns may be problematic.

The first intestinal discharge from newborns is meconium, which is a viscous, dark green substance composed of intestinal secretions, desquamated squamous cells, lanugo hair, bile pigments, and blood. Meconium also contains pancreatic enzymes, free fatty acids, porphyrins, interleukin-8, and phospholipase A_2 primary bile acids with a small quantity of secondary bile acids. Meconium begins to form during the second trimester and continues to accumulate until birth; drugs taken by the mother can be detected in the meconium of the newborn [114].

The disposition of drug in meconium is not well understood. The proposed mechanism is that the fetus excretes drug into bile and amniotic fluid. Drug accumulates in meconium by direct deposition from bile or through swallowing of amniotic fluid [114,115]. The first evidence of meconium in the fetal intestine appears at approximately the 10th−12th week of gestation; meconium slowly moves into the colon by the 16th week of gestation [116]. Therefore the presence of drugs in meconium has been proposed to be indicative of in utero drug exposure up to ~5 months before birth—a longer historical measure than is possible by urinalysis [114].

The collection of meconium is noninvasive, making sample collection easy; however, meconium testing does have some limitations. Meconium is usually passed by full-term newborns within 24−48 hours, after which transition to stool. Infants with low birth weight (<1000 g) have been shown to pass their first meconium at a median age of 3 days. Thus meconium collection is missed because of delayed passage, and meconium may not be available soon after birth for early detection of intrauterine drug exposure.

In the clinical laboratory, meconium is an unfamiliar matrix; it is a sticky material that is more difficult to work with than urine. Meconium drug screening has been adapted to various analytical techniques, including radioimmunoassay, enzyme immunoassay, and fluorescence polarization immunoassay. Urine drugs-of-abuse screening assays frequently use meconium extracts and therefore must be investigated for possible effects of matrix on accuracy, precision, and assay linearity. However, some other immunoassay screening methods have been used. As with any immunoassay-based drug screen, confirmation by MS is critical. Confirmation assays for meconium are more difficult than those for urine. Recovery of drugs from meconium is sometimes low (30%−50%). A variety of GC−MS, LC−MS, and LC−MS/MS methods and their advantages and disadvantages have been described elsewhere [77,117,118].

Umbilical cord tissue

Alternatively, umbilical cord testing has been proposed as an alternative sample type to detect intrauterine drug exposure. It has some potential advantages over meconium in that sample quantity is not an issue, it is easier to collect in a single collection, and you do not have the potential of missing it, since meconium can be passed in utero. However, many questions also still remain to be answered about the disposition of drugs in this matrix. In a study comparing the detection of drugs in umbilical cord tissue and meconium for the confirmation of in utero drug exposures, the authors showed that cord tissue was less sensitive for the detection of five of the six drug classes [119]. Other studies [120] showed that drugs detected in paired meconium and umbilical cord tissue samples are often discordant; therefore the tests are not equivalent. However, one test is not always superior. Therefore it is a valid alternative sample type being utilized.

Oral fluid

Analysis of saliva for drugs was first used almost 30 years ago for the purpose of therapeutic drug monitoring [121]. It has since been evaluated for use in toxicology, with recognition of its advantages over other biological matrices. Most studies on saliva in humans use whole saliva. The term “oral fluid” is preferred for the specimen collected from the mouth. Oral fluid is a complex fluid consisting not only of secretions from the three major pairs

of salivary glands (parotid, submandibular, and sublingual), but also of secretions from the minor glands (labial, buccal, and palatal), bacteria, sloughed epithelial cells, gingival fluid, food debris, and other particulate matter [122]. The concentration of drug from each secretion and the relative contributions of the various glands to the final fluid may vary [123].

Several advantages are associated with monitoring oral fluid as contrasted with monitoring plasma or serum concentrations [124]. Collection of oral fluid is considered to be a noninvasive procedure, and some of the risks associated with the drawing of blood are avoided. Furthermore, for the patient, fear, anxiety, and discomfort that may accompany the drawing of blood are diminished. A disadvantage is the small volume of sample collected, and the window of detection for drugs is shorter than for urine. The problem of small sample size can be overcome by using methods that simultaneously detect multiple drug groups [125,126].

In recent years, great interest has been expressed in the use of oral fluid testing for monitoring the compliance of individuals on drug maintenance programs, and for workplace drug testing. Low concentrations of drugs and metabolites necessitate sensitive screening methods, which typically are immunoassays [126]. Again, the low concentrations of drugs have necessitated that confirmatory methods be equally as sensitive. Many confirmatory methods have been developed for oral fluid testing of abused drugs, including GC−MS, LC−MS, and LC−MS/MS.

Hair

For more than 30 years, hair has been analyzed for toxic metals, including lead, arsenic, and mercury. At first, the examination of hair for organic substances, specifically drugs, was not possible, because the analytical methods were not sensitive enough [127]. Baumgartner et al. [128] published the first report on the detection of morphine in the hair of heroin abusers using RIA. Since that time, interest in analysis of hair for the purpose of detecting drug use has increased.

It is generally accepted that drugs enter hair by at least three mechanisms: (1) from blood that supplies the growing hair follicle; (2) through sweat and sebum; and (3) via the external environment [127]. The exact mechanism by which chemicals are incorporated into hair is not known. Factors that may affect how efficiently drugs are incorporated into hair are not well established but may include rate of hair growth, anatomic location of hair, hair color (melanin content), and hair texture (thick or fine, and porous or not); these are determined by genetic factors and by the effects of various hair treatments [127]. For example, in-vitro substantially higher binding was found with hair from black men than with hair from blond white men, suggesting that melanin pigment plays an important role in drug binding. Studies have demonstrated that, after the same dosage is given, dark hair incorporates much more of the drug than is incorporated by blond hair [129]. This may lead to bias in hair testing for drugs of abuse and discussions about possible genetic variability of drug deposition in hair and is still under evaluation.

Hair is advantageous as a biological specimen, because it is easily obtained, with less embarrassment; it is not easily altered or manipulated to avoid drug detection. Once deposited in hair, drugs are very stable, and analysis can be performed even after centuries [130]. Hair also differs from other human materials used for toxicological analysis, such as blood or urine, because of its substantially longer detection window (months to years). Hair grows at a relatively constant rate. The average rate of hair growth is usually stated to be 0.44 mm/day.

Drugs, when deposited in hair, are generally present in relatively low concentrations (pg/ng/μg/mg); thus sensitive analytical techniques are required for detection. Immunoassay procedures have been modified for use with hair [127]. Although GC−MS is generally the method of choice, various GC−MS/MS or LC−MS/MS methods may be used for targeted analysis. As mentioned, external exposure to drugs causes them to be detected in hair. Externally deposited substances easily contaminate the hair because of its high surface-to-volume ratio. Substances deposited in hair from the environment are loosely bound to the surface of the hair, and thus are removed by appropriate decontamination procedures. These usually involve a washing step [127,131]. It is fundamental to be able to distinguish between passive exposure (environmental contamination) and actual use.

Metals

Elemental toxicity is still a concern today despite having a better awareness and understanding of the dangers. On a daily basis, people are exposed to various elements through the environment, their occupation, and/or the consumption of food or water. While some elements are absolutely essential for life (e.g., sodium, potassium, and magnesium), others may be essential but only required in very low concentrations (e.g., chromium), and still others have absolutely no biological role at all (e.g., lead). Toxic heavy metals (e.g., lead, mercury, and arsenic) are individual metals or metal compounds that can negatively affect people's health. In very small amounts, many of the elements may not cause any toxicity, but, when they are present in larger amounts, they can become toxic.

Diagnosing metal toxicity

Laboratory tests play a key role in confirming the diagnosis of elemental toxicity. Demonstration of all of the

following is necessary to make a definitive diagnosis including (1) an identifiable source for the elemental exposure; (2) the patient has clinical signs and symptoms of toxicity consistent with the identified element; and (3) an abnormal concentration of the element is found in the appropriate test matrix (e.g., blood, serum, or urine). Since elements are ubiquitous in the environment and the laboratory result plays a key role in the diagnosis, it is critical that the preanalytical variables are controlled to minimize sample contamination that can lead to an incorrect diagnosis. For example, most of the common blood collection tubes used have rubber stoppers made of aluminum silicate, which could lead to falsely elevated (abnormal) aluminum results if a sample is collected in these tubes. As a result, special metal-free collection tubes, collection techniques, and processing/handling of serum samples are required to prevent misinterpretation or misdiagnosis of element toxicity.

Lead

Lead is a naturally occurring blue–white metal, which is ubiquitous in the environment and has no known biological function in humans. Lead is used in the production of batteries, ammunition, and pipes. Common sources of exposure include older household paints (prior to 1972), ceramic products, leaded gasoline (prior to 1978), soil (especially near freeways), and water transported through lead pipes or lead soldered pipes. Due to the serious adverse effects of low levels of lead in children, the CDC and American Academy of Pediatrics recommend lead screening in all children who are Medicaid-eligible or enrolled and in those children with identified higher risks (e.g., live in older housing and have a sibling with elevated levels). The CDC reference blood lead level for adults and children is <5 μg/dL. Since children are lower to the ground where lead-contaminated dust accumulates and they engage in hand-to-mouth activities, they ingest and inhale more lead-contaminated dust. While ingested lead is poorly absorbed ($<10\%$) in adults, up to 50% may be absorbed in children especially those who are iron- or calcium-deficient [117]. Once absorbed, more lead also crosses the blood–brain barrier in children and has effects on their developing nervous system.

Clinical signs and symptoms of lead toxicity can include severe abdominal pain, vomiting, headaches, poor concentration, weight loss, ataxia, seizures, and death. One of the mechanisms of toxicity for lead is that it inhibits several enzymes that are necessary for the synthesis of heme from porphyrin including amino levulinic acid dehydratase. Inhibition of this enzyme leads to an accumulation of protoporphyrin in erythrocytes and anemia. To detect or confirm exposure to lead, a venous collected whole blood sample is the preferred specimen type for lead testing. The Occupational Safety Health Administration (OSHA) also requires monitoring of zinc protoporphyrin along with the whole blood lead level to monitor occupational exposure. The type of treatment depends on the patient's blood lead level, but asymptomatic patients usually only requiring separation from the source of exposure where chelation therapy using oral dimercaprol or D-penicillamine can be used for the more severe lead poisonings/toxicity.

Arsenic

Arsenic has a long history as a poison used to eliminate undesirable acquaintances. It is listed as the No. 1 toxicant on the 2015 Agency for Toxic Substances and Disease Registry Substance Priority List of Hazardous Substances based on a combination of frequency, toxicity, and potential for human exposure (http://www.atsdr.cdc.gov/SPL/index.html, accessed November 15, 2018). Arsenic exists in numerous toxic (inorganic) and nontoxic (organic) forms. Nontoxic forms of arsenic (arsenobetaine and arsenocholine) are present in many foods especially shellfish and predator fish (e.g., cod and haddock). The toxic forms include arsenite (As_{3+}), the more toxic arsenate (As_{5+}), and the less toxic metabolites; monomethylarsonic acid; and dimethylarsinic acid. The largest source of exposure is from pesticides, but arsenic is also used in pharmaceuticals, the glass and ceramic industry, and metallurgy.

Acute arsenic toxicity can present quickly with abdominal pain, vomiting, diarrhea, muscle cramps, hypertension, and tachycardia. Several weeks after the exposure, transverse white bands (mee's lines) on the fingernails may be seen. Chronic arsenic toxicity typically presents with dermatological (hyperpigmentation), neurological (decreased muscle strength in lower extremities), hematological (anemia and leukopenia), and GI symptoms. Since blood concentrations of arsenic are only elevated for only a short time after administration, urine measurements are the preferred matrix to screen for arsenic poisoning. However, if the total arsenic concentration is elevated, a speciation or fractionation test must be performed to determine if the toxic (inorganic) forms or nontoxic (organic) forms of arsenic are present. Therefore it is recommended that patients abstain from eating seafood several days before collecting a urine sample for arsenic testing. Treatment consists of removal from the source of exposure, supportive measures, and chelation therapy (e.g., British antilewisite).

Cadmium

Cadmium is ubiquitous and found in all rocks and soil. It is used in batteries, electroplating, as pigments in paints,

and as a stabilizer in plastic production. Exposure to cadmium can occur through eating food, drinking water, smoking tobacco, or various occupational exposures. Symptoms of cadmium poisoning include nausea, vomiting, abdominal pain, and diarrhea. In a similar fashion to arsenic, mercury, and lead, cadmium also damages the kidneys leading to glucosuria and proteinuria, which slowly over time can lead to renal and hepatic failure. Typically, blood cadmium concentrations can be used to monitor acute toxicity, while urine levels can be used to assess body burden. For occupational monitoring, the OSHA recommends a panel that includes cadmium in blood and urine, as well as beta-2-microglobulin in urine. Collection of urine samples using a rubber catheter or use of colored plastic urine collection containers should be avoided, since it can falsely elevate the test results due to the cadmium in the rubber and pigments in the plastic. Therapy includes removal of the patient from the source and supportive care. Chelation therapy is contraindicated, since it exposes the kidney to large quantities of the nephrotoxic cadmium.

Chromium

Chromium is a naturally occurring element found in rocks, animals, plants, and soil. Chromium is present in several different forms [e.g., chromium(0), chromium(III), and chromium(VI)]. It is used extensively in the manufacturing of stainless steel, chrome plating, leather tanning, dyes for printing, and as a cleaning solution. It is an essential element, and a deficiency of chromium is associated with impaired glucose tolerance and diabetes. Symptoms associated with chromium toxicity vary based upon route of exposure and dose, and may include dermatitis, impairment of pulmonary function, gastroenteritis, hepatic necrosis, bleeding, and acute tubular necrosis. More recently, chromium toxicity due to the release of chromium ions during the normal and abnormal wear of metal-on-metal (MOM) prosthetic implants has gotten attention. Orthopedic implants are created from various metal alloys with the most common being cobalt and chromium. As a result, the FDA recommended monitoring MOM wear using whole blood measurement of chromium to minimize preanalytical variables and contamination in symptomatic patients (http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/ImplantsandProsthetics/MetalonMetalHipImplants/ucm331971.htm, accessed November 15, 2018). Urine chromium levels can also be used to monitor other short-term exposure.

Gadolinium

Gadolinium (Gd) is an element that is commonly found in image contrast agents used during magnetic resonance imaging (MRI) studies. These gadolinium-based contrast agents (GBCAs) increase the clinical utility and sensitivity of these exams, providing enhanced medical care by improving the effectiveness of the imaging studies [132]. An estimated 30-million Gd-enhanced MRIs are conducted around the world each year [133]. Recently these agents have come under scrutiny, and, in 2017 the FDA released several safety announcements through its MedWatch system regarding the use of GBCAs. The new warning concerns the fact that Gd can be retained in patients bodies (e.g., the brain and other tissues) for months to years after receiving these drugs. The December 2017 FDA announcement also noted that there are differences in the two types of GBCAs (e.g., linear and macrocyclic) used. Of the two, linear agents result in more retention of Gd in patients, and retention for a longer period of time than macrocyclic GBCAs. Regardless of the new warnings, the FDA concluded that the benefit of the approved GBCAs continues to outweigh any potential risk—the European Medicines Agency reached a similar conclusion. Although ongoing studies continue to examine if there are any potential adverse effects of the long-term retention of Gd, the only known adverse health effect to date related to Gd retention is a rare condition called nephrogenic systemic fibrosis (NSF) [134]. NSF is a relatively uncommon condition where fibrous plaques develop in the dermis and, often, in deeper connective tissues, and is characterized by thickening of the skin [134]. Reported cases of NSF have occurred almost exclusively in patients with severe renal disease, and almost all have been associated with prior use of GBCAs.

Measurement of Gd is not routinely conducted nor formally recommended by the FDA, or any professional practice guidelines. However, laboratories using inductively coupled plasma−mass spectrometry (ICP−MS) in the assessment of other elements should check for Gd in patients recently administered GBCAs, since it can suppress the measurement of other elements due to space-charge effects, resulting in a loss of accuracy/recovery for the element of interest [134]. Serum and urine Gd testing is available, but only confirms past exposure to Gd. Currently, there is no definitive or diagnostic test for Gd toxicity.

Mercury

Hg is widely found in the environment and occurs both naturally and as the result of industrial processes, with the single largest source of Hg being its natural out-gassing from granite rock. Metallic mercury is a silvery, slightly volatile liquid at room temperature. Mercury can also combine with other elements, such as chlorine, sulfur, or oxygen, to form inorganic mercury compounds or "salts." It can also combine with carbon to make organic mercury compounds. The most common one, methylmercury, is

produced mainly by microscopic organisms in the water and soil. Previously, mercury was used in thermometers and barometers, but those are being phased out due to concerns about its toxicity. Mercury has also been used in dental amalgams, lighting as mercury vapor lamps, in the pulp and paper industry as a whitener, as a catalyst in the synthesis of plastics, and as a potent fungicide in antifouling and latex paints.

Symptoms of mercury toxicity depend on the type of exposure (inhalation, ingestion, or dermal) and on the chemical form of mercury. Symptoms of acute mercury poisoning include dermatologic (blue−black pigmentation), neurologic (headaches, confusion, and ataxia), GI (metallic taste, nausea, and vomiting), and hematologic (anemia and thrombocytopenia). Laboratory assessment of mercury exposure can be done using whole blood, urine, or hair. Urine can specifically be used to follow the effectiveness of chelation therapy and is thought to indicate better the mercury burden of the kidneys. Treatment includes removal from the source of exposure, supportive care, decontamination, and chelation therapy.

Analytical techniques used to measure metals in biological fluid include atomic absorption spectrometry, ICP−MS, and LC−ICP−MS. These techniques are specific, sensitive, and provide the clinical laboratory with the capability to measure a broad array of metals at clinically significant concentrations.

Tricyclic antidepressants

Tricyclic antidepressants (TCAs) are frequently prescribed for the treatment of depression and neuropathic pain. The TCAs have been largely supplanted by the newer, less toxic selective serotonin reuptake inhibitors and other atypical agents for depression, which now are accepted broadly as drugs of first choice, particularly for medically ill or potentially suicidal patients and for the elderly and the young.

Cardiovascular effects, the most serious manifestation of TCA overdose, account for the majority of fatalities. Several mechanisms may contribute to cardiovascular toxicity [135]:

1. Anticholinergic effects and inhibition of neuronal reuptake of catecholamines result in tachycardia and potential for early hypertension.
2. Peripheral alpha-adrenergic blockade causes vasodilation and contributes to hypotension.
3. Cardiac fast sodium channel blockade and potassium efflux blockade, resulting in QRS and QT prolongation, respectively, and risk for decreased cardiac output and ventricular dysrhythmias.
4. Metabolic and/or respiratory acidosis that may further contribute to inhibiting the cardiac fast sodium channel.

TCAs are measured by chromatographic or immunoassay methods. Immunoassays are rapid and relatively easy to perform, but may be subject to interference by other drugs, such as chlorpromazine, thioridazine, cyclobenzaprine, and diphenhydramine [136−138], and are not able to necessarily identify which TCA is being quantitated. In cases of overdose, qualitative identification (serum or urine) is sufficient, because the severity of intoxication is more reliably indicated by an increase in the QRS interval than by the serum concentration.

Agents related to cholinergic syndrome

Agents inducing cholinergic syndrome are diverse, but the commonality is their ability to produce uncontrolled acetylcholine transmission through inactivation of cholinesterase enzymes or direct stimulation of acetylcholine receptors. The duration of acetylcholine action is controlled by AChE and butyrylcholinesterase (pseudocholinesterase). AChE is found in red blood cells, nervous tissue, and skeletal muscle. Butyrylcholinesterase is found in plasma, liver, heart, pancreas, and brain.

Pesticides (organophosphate and carbamate compounds)

Organophosphate (e.g., malathion, parathion, diazinon, and Dursban) and carbamate (e.g., Sevin and Furadan) insecticides, as well as military nerve agents [e.g., sarin (GB), soman (GD), tabun (GA), and VX], exert their toxicity by inhibiting the action of AChE, thereby causing a pronounced cholinergic response [5]. Excess synaptic acetylcholine stimulates muscarinic receptors (peripheral and CNS) and stimulates but then depresses or paralyzes nicotinic receptors. Activation of peripheral muscarinic receptors causes signs and symptoms described by the mnemonics SLUDGE or DUMB BELLS (see “Toxidromes” section). Whereas, CNS neurotoxic effects include (1) restlessness; (2) agitation; (3) lethargy; (4) confusion; (5) slurred speech; (6) seizures; (7) coma; (8) cardiorespiratory depression; and (9) death. Stimulation or paralysis of nicotinic receptors at the neuromuscular junction causes muscle fasciculation, cramping, weakness, and respiratory muscle paralysis; stimulation of nicotinic receptors at sympathetic ganglia results in a sympathomimetic syndrome that includes hypertension, tachycardia, pallor, and mydriasis. The actual signs and symptoms observed with these toxins depend on the balance of muscarinic and nicotinic receptor activation.

Specific therapy for organophosphate and carbamate insecticide poisoning includes administration of atropine to block the muscarinic (but not nicotinic) actions of acetylcholine. In addition, pralidoxime is given to reactivate cholinesterase. Pralidoxime binds to the cholinesterase catalytic site and, via nucleophilic attack by its oxime

group, dephosphorylates or decarbamylates the serine group. Administration of pralidoxime may not be necessary in cases of carbamate insecticide poisoning, because carbamylated cholinesterase spontaneously reactivates within a few hours.

Diagnosis of organophosphate and carbamate toxicity depends mainly on exposure history, physical presentation, clinical suspicion, and laboratory support. Cholinesterase activity is measured to assess exposure and to monitor reactivation during treatment. AChE and butyrylcholinesterase enzyme activity are typically monitored using Ellman colorimetric or Michel electrometric analyses using whole blood, plasma, serum, and dried blood spots [139,140].

References

[1] T.B. Erickson, T.M. Thompson, J.J. Lu, The approach to the patient with an unknown overdose, Emerg. Med. Clin. North. Am. 25 (2007) 249–281. abstract vii.

[2] A. Nice, J.B. Leikin, A. Maturen, L.J. Madsen-Konczyk, M. Zell, D.O. Hryhorczuk, Toxidrome recognition to improve efficiency of emergency urine drug screens, Ann. Emerg. Med. 17 (1988) 676–680.

[3] G.A. McMillin, M.H. Slawson, S.J. Marin, K.L. Johnson-Davis, Demystifying analytical approaches for urine drug testing to evaluate medication adherence in chronic pain management, J. Pain. Palliat. Care Pharmacother. 27 (2013) 322–339.

[4] D.C. Keyes, R.C. Dart, Initial diagnosis and treatment of the posnoned patient, in: R.C. Dart (Ed.), Medical Toxicology, Lippincott Williams & Wilkins, Philadelphia, PA, 2004, pp. 21–31.

[5] R.F. Clarke, Insecticides: organic phosphorus compounds and carbamates, in: N.E. Flomenbaum, L.R. Goldfrank, R.S. Hoffman, M.A. Howland, N.A. Lewin, L.S. Nelson (Eds.), Goldfrank's Toxicologic Emergencies, eighth ed, McGraw-Hill, New York, 2006, pp. 1497–1512.

[6] K. Ishihara, H.M. Szerlip, Anion gap acidosis, Semin. Nephrol. 18 (1998) 83–97.

[7] L.J. Langman, L.K. Bechtel, B.M. Meier, C. Holstege, Chapter 41: Clinical toxicology, in: N. Rifai, A.R. Horvath, C.T. Wittwer (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, sixth ed, Elsevier, St Louis, MO, 2018, pp. 832–887.

[8] J. Krahn, A. Khajuria, Osmolality gaps: diagnostic accuracy and long-term variability, Clin. Chem. 52 (2006) 737–739.

[9] C. Tomaszewski, Carbon monoxide, in: G. Flomenbaum, H. Hoffman, N. Lewin (Eds.), Goldfrank's Toxicologic Emergencies, eighth ed, McGraw Hill, New York, 2006, pp. 1689–1704.

[10] K.R. Olson, Poisoning, in: K.R. Olson (Ed.), Current Medical Diagnosis & Treatment, sixth ed, McGraw-Hill Companies, Inc, 2015. Available from: http://accessmedicine.mhmedical.com/content.aspx?bookid = 391&Sectionid = 42069858.

[11] D. Price, Methemoglobin inducers, in: G. Flomenbaum, H. Hoffman, N. Lewin (Eds.), Goldfrank's Toxicologic Emergencies, eighth ed, McGraw Hill, New York, 2006, pp. 1734–1745.

[12] P.D. Blanc, Chapter 103: Methemoglobinemia, in: K.R. Olson (Ed.), Poisoning & Drug Overdose, sixth ed, The McGraw-Hill Companies, New York, 2012.

[13] K.A. Evelyn, H.T. Malloy, Microdetermination of oxyhemoblobin, methemoglobin, and sulfhemoglobin in a single sample of blood, J. Biol. Chem. 126 (1938) 655–662.

[14] V. Schulz, Clinical pharmacokinetics of nitroprusside, cyanide, thiosulphate and thiocyanate, Clin. Pharmacokinet. 9 (1984) 239–251.

[15] D.L. Morgan, D.J. Borys, Chapter 47: Poisoning, in: C.K. Stone, R.L. Humphries (Eds.), Current Diagnosis & Treatment Emergency Medicine, seventh ed, The McGraw-Hill Companies, New York, 2011.

[16] R.C. Baselt, in: R.C. Baselt (Ed.), Dispositition of Toxic Drugs and Chemical in Man, tenth ed, Biomedical Publications, Seal Beach, CA, 2014.

[17] B. Levine, Y.H. Caplan, Chapter 2: Pharmacology and toxicology of alcohol, in: Y.H. Caplan, B.A. Goldberger (Eds.), Garriott's Medicolegal Aspects of Alcohol, sixth ed, Lawyers & Judges Publishing Company, Inc, Tucson, AZ, 2015, pp. 25–48.

[18] K.M. Dubowski, Absorption, distribution and elimination of alcohol: highway safety aspects, J. Stud. Alcohol. Suppl. 10 (1985) 98–108.

[19] M.L. Sivilotti, Ethanol, isoporpanol, and methanol, in: R.C. Dart (Ed.), Medical Toxicology, Lippincott Williams & Wilkins, Philadelphia, PA, 2004, pp. 1211–1223.

[20] M.J. Matteucci, Chapter 89: Isopropyl alcohol, in: K.R. Olson (Ed.), Poisoning & Drug Overdose, sixth ed, The McGraw-Hill Companies, New York, 2012.

[21] H.A. Jolliff, M.L. Sivilotti, Ethylene glycol, in: R.C. Dart (Ed.), Medical Toxicology, Lippincott Williams & Wilkins, Philadelphia, PA, 2004, pp. 1223–1230.

[22] J.P. Payne, D.W. Hill, D.G. Wood, Distribution of ethanol between plasma and erythrocytes in whole blood, Nature 217 (1968) 963–964.

[23] C.L. Winek, M. Carfagna, Comparison of plasma, serum, and whole blood ethanol concentrations, J. Anal. Toxicol. 11 (1987) 267–268.

[24] R.H. Gadsden Jr., E.H. Taylor, S.J. Steindel, Ethanol in biological fluids by enzymic analysis, in: C. Frings, W. Faulkner (Eds.), Selected Methods of Emergency Toxicology, AACC Press, Washington, DC, 1986, pp. 63–65.

[25] R.A. Middleberg, L.J. Langman, Chapter 16: Clinical aspects of alcohol testing, in: Y.H. Caplan, B.A. Goldberger (Eds.), Garriott's Medicolegal Aspects of Alcohol, sixth ed, Lawyers & Judges Publishing Company Inc, Tucson, AZ, 2015, pp. 435–448.

[26] H. Dahl, N. Stephanson, O. Beck, A. Helander, Comparison of urinary excretion characteristics of ethanol and ethyl glucuronide, J. Anal. Toxicol. 26 (2002) 201–204.

[27] A. Helander, O. Beck, Ethyl sulfate: a metabolite of ethanol in humans and a potential biomarker of acute alcohol intake, J. Anal. Toxicol. 29 (2005) 270–274.

[28] G. Viel, R. Boscolo-Berto, G. Cecchetto, P. Fais, A. Nalesso, S.D. Ferrara, Phosphatidylethanol in blood as a marker of chronic alcohol use: a systematic review and meta-analysis, Int. J. Mol. Sci. 13 (2012) 14788–14812.

[29] A. Helander, I. Olsson, H. Dahl, Postcollection synthesis of ethyl glucuronide by bacteria in urine may cause false identification of alcohol consumption, Clin. Chem. 53 (2007) 1855–1857.

[30] Substance Abuse and Mental Health Services Administration, the Role of Biomarkers in the Treatment of Alcohol Use Disorders, 2012 Revision Vol, HHS Publication No. (SMA) 12-4686: SAMHSA Advisory, 2012.
[31] G.M. Reisfield, B.A. Goldberger, B.O. Crews, A.J. Pesce, G.R. Wilson, S.A. Teitelbaum, et al., Ethyl glucuronide, ethyl sulfate, and ethanol in urine after sustained exposure to an ethanol-based hand sanitizer, J. Anal. Toxicol. 35 (2011) 85–91.
[32] G.M. Reisfield, B.A. Goldberger, A.J. Pesce, B.O. Crews, G.R. Wilson, S.A. Teitelbaum, et al., Ethyl glucuronide, ethyl sulfate, and ethanol in urine after intensive exposure to high ethanol content mouthwash, J. Anal. Toxicol. 35 (2011) 264–268.
[33] R.B. Palmer, A review of the use of ethyl glucuronide as a marker for ethanol consumption in forensic and clinical medicine, Semin. Diagn. Pathol. 26 (2009) 18–27.
[34] T. Grosser, E. Smyth, G. FitzGerald, Pharmacotherapy of inflammation, fever, pain, and gout, in: L.L. Brunton, R. Hilal-Dandan, B.C. Knollmann (Eds.), Goodman & Gilman's: The Pharmacological Basis of Therapeutics, 13th ed, McGraw-Hill Education, New York, 2017.
[35] J.R. Mitchell, S.S. Thorgeirsson, W.Z. Potter, D.J. Jollow, H. Keiser, Acetaminophen-induced hepatic injury: protective role of glutathione in man and rationale for therapy, Clin. Pharmacol. Ther. 16 (1974) 676–684.
[36] H. Perry, M.W. Shannon, Acetaminophen, in: M. Hl, W. Sm, F. Wj (Eds.), Clinical Management of Poisoning and Drug Overdose, third ed, WB Saunders Co, Philadelphia, PA, 1998, pp. 664–674.
[37] R.J. Flanagan, T.G. Mant, Coma and metabolic acidosis early in severe acute paracetamol poisoning, Hum. Toxicol. 5 (1986) 179–182.
[38] B.H. Rumack, H. Matthew, Acetaminophen poisoning and toxicity, Pediatrics 55 (1975) 871–876.
[39] C. Dale, A.A. Aulaqi, J. Baker, R.C. Hobbs, M.E. Tan, C. Tovey, et al., Assessment of a point-of-care test for paracetamol and salicylate in blood, QJM 98 (2005) 113–118.
[40] A.K. Done, Salicylate intoxication. Significance of measurements of salicylate in blood in cases of acute ingestion, Pediatrics 26 (1960) 800–807.
[41] P. Trinder, Rapid determination of salicylate in biological fluids, Biochem. J. 57 (1954) 301–303.
[42] American College of Obstetricians and Gynaecologists, Committee opinion no. 422: at-risk drinking and illicit drug use: ethical issues in obstetric and gynecologic practice, Obstet. Gynecol. 112 (2008) 1449–1460.
[43] B.K. Logan, Methamphetamine—effects on human performance and behavior, Forensic Sci. Rev. 14 (2002) 133–151.
[44] C.L. Hornbeck, R.J. Czarny, Retrospective analysis of some l-methamphetamine/l-amphetamine urine data, J. Anal. Toxicol. 17 (1993) 23–25.
[45] C. Lüscher, Chapter 32: Drugs of abuse, in: B.G. Katzung (Ed.), Basic & Clinical Pharmacology, eleventh ed, McGraw-Hill Companies, Inc, 2009. Available from: http://www.accessmedicine.com/content.aspx?aID = 4519820.
[46] K.A. Moore, Chapter 18: Amphetamines/sympathomimetic amines, in: B. Levine (Ed.), Principles of Forensic Toxicology, third ed, AACC Press, Washington, DC, 2010, pp. 305–322.
[47] B.K. Logan, F.A. Cooper, 3,4-Methylenedioxymethamphetamine—effects on human performance and behavior, Forensic Sci. Rev. 15 (2003) 11–28.
[48] M.R. Meyer, F.T. Peters, H.H. Maurer, The role of human hepatic cytochrome P450 isozymes in the metabolism of racemic 3,4-methylenedioxy-methamphetamine and its enantiomers, Drug. Metab. Dispos. 36 (2008) 2345–2354.
[49] S.S. Johansen, A.C. Hansen, I.B. Muller, J.B. Lundemose, M.B. Franzmann, Three fatal cases of pma and pmma poisoning in denmark, J. Anal. Toxicol. 27 (2003) 253–256.
[50] M. Rohanova, M. Balikova, Studies on distribution and metabolism of para-methoxymethamphetamine (pmma) in rats after subcutaneous administration, Toxicology 259 (2009) 61–68.
[51] L.C. Daws, R.J. Irvine, P.D. Callaghan, N.P. Toop, J.M. White, F. Bochner, Differential behavioural and neurochemical effects of para-methoxyamphetamine and 3,4-methylenedioxymethamphetamine in the rat, Prog. Neuropsychopharmacol. Biol. Psychiatry 24 (2000) 955–977.
[52] C.L. German, A.E. Fleckenstein, G.R. Hanson, Bath salts and synthetic cathinones: an emerging designer drug phenomenon, Life Sci. 97 (2014) 2–8.
[53] A.G. Verstraete, F.V. Heyden, Comparison of the sensitivity and specificity of six immunoassays for the detection of amphetamines in urine, J. Anal. Toxicol. 29 (2005) 359–364.
[54] S.J. Mihic, J. Mayfield, R.A. Harris, Hypnotics and sedatives, in: L. L. Brunton, R. Hilal-Dandan, B.C. Knollmann (Eds.), Goodman & Gilman's: The Pharmacological Basis of Therapeutics, thirteenth ed, McGraw-Hill Education, New York, 2017.
[55] B. Levine, Chapter 14: Miscellaneous central nervous system depressants, in: B. Levine (Ed.), Principles of Forensic Toxicology, second ed, AACC Press, Washington, DC, 2010, pp. 215–224.
[56] O.H. Drummer, Benzodiazepines—effects on human performance and behavior, Forensic Sci. Rev. 14 (2002) 1–14.
[57] K.B. Green, D.S. Isenschmid, Medical review officer interpretation of urine drug testing results, Forensic Sci. Rev. 7 (1995) 41–60.
[58] M.A. Huestis, Chapter 17: Cannabis, in: B. Levine (Ed.), Principles of Forensic Toxicology, second ed, AACC Press, Washington, DC, 2010, pp. 269–303.
[59] M. McGuigan, Cannabinoids, in: L.R. Goldfrank, N.E. Flomenbaum, R.S. Hoffman, M.A. Howland, N.A. Lewin, L.S. Nelson (Eds.), Goldfrank's Toxicologic Emergencies, eight ed, McGraw-Hill, New York, 2006, pp. 1212–1230.
[60] D.M. Schwope, E.L. Karschner, D.A. Gorelick, M.A. Huestis, Identification of recent cannabis use: whole-blood and plasma free and glucuronidated cannabinoid pharmacokinetics following controlled smoked cannabis administration, Clin. Chem. 57 (2011) 1406–1414.
[61] Product information: Epidiolex(r) oral solution, cannabidiol oral solution. Vol. Greenwich Biosciences Inc (per FDA), Carlsbad, CA, 2018.
[62] S.J. Mule, P. Lomax, S.J. Gross, Active and realistic passive marijuana exposure tested by three immunoassays and GC/MS in urine, J. Anal. Toxicol. 12 (1988) 113–116.
[63] B. Law, P.A. Mason, A.C. Moffat, L.J. King, V. Marks, Passive inhalation of cannabis smoke, J. Pharm. Pharmacol. 36 (1984) 578–581.
[64] M.L. Smith, A.J. Barnes, M.A. Huestis, Identifying new cannabis use with urine creatinine-normalized thccooh concentrations and time intervals between specimen collections, J. Anal. Toxicol. 33 (2009) 185–189.

[65] A.D. Fraser, D. Worth, Urinary excretion profiles of 11-nor-9-carboxy-delta9-tetrahydrocannabinol. Study III. A delta-9-THC-COOH to creatinine ratio study, Forensic Sci. Int. 137 (2003) 196–202.
[66] E.J. Cone, R. Lange, W.D. Darwin, In vivo adulteration: excess fluid ingestion causes false-negative marijuana and cocaine urine test results, J. Anal. Toxicol. 22 (1998) 460–473.
[67] E.W. Schwilke, R.G. Gullberg, W.D. Darwin, C.N. Chiang, J.L. Cadet, D.A. Gorelick, et al., Differentiating new cannabis use from residual urinary cannabinoid excretion in chronic, daily cannabis users, Addiction 106 (2011) 499–506.
[68] O. Aizpurua-Olaizola, I. Zarandona, L. Ortiz, P. Navarro, N. Etxebarria, A. Usobiaga, Simultaneous quantification of major cannabinoids and metabolites in human urine and plasma by HPLC-MS/MS and enzyme-alkaline hydrolysis, Drug. Test. Anal. 9 (2017) 626–633.
[69] M.M. Bergamaschi, A. Barnes, R.H. Queiroz, Y.L. Hurd, M.A. Huestis, Impact of enzymatic and alkaline hydrolysis on CBD concentration in urine, Anal. Bioanal. Chem. 405 (2013) 4679–4689.
[70] M.A. ElSohly, H. deWit, S.R. Wachtel, S. Feng, T.P. Murphy, Delta9-tetrahydrocannabivarin as a marker for the ingestion of marijuana versus marinol: results of a clinical study, J. Anal. Toxicol. 25 (2001) 565–571.
[71] E. Hidvegi, G.P. Somogyi, Detection of cannabigerol and its presumptive metabolite in human urine after cannabis consumption, Pharmazie 65 (2010) 408–411.
[72] P.M. Kemp, I.K. Abukhalaf, J.E. Manno, B.R. Manno, D.D. Alford, G.A. Abusada, Cannabinoids in humans. I. Analysis of delta 9-tetrahydrocannabinol and six metabolites in plasma and urine using GC-MS, J. Anal. Toxicol. 19 (1995) 285–291.
[73] K.B. Scheidweiler, A.J. Barnes, Quantification of eight cannabinoids including cannabidiol in human urine via liquid chromatography tandem mass spectrometry, Methods Mol. Biol. 1872 (2019) 11–22.
[74] D.S. Isenschmid, Chapter 16: Cocaine, in: B. Levine (Ed.), Principles of Forensic Toxicology, second ed, AACC Press, Washington, DC, 2010, pp. 245–268.
[75] D.S. Isenschmid, Cocaine—effects on human performance and behavior, Forensic Sci. Rev. 14 (2002) 61–100.
[76] R.A. Dean, C.D. Christian, R.H. Sample, W.F. Bosron, Human liver cocaine esterases: ethanol-mediated formation of ethylcocaine, FASEB J. 5 (1991) 2735–2739.
[77] C. Moore, A. Negrusz, D. Lewis, Determination of drugs of abuse in meconium, J. Chromatogr. 713 (1998) 137–146.
[78] B.M. Lester, M. ElSohly, L.L. Wright, V.L. Smeriglio, J. Verter, C.R. Bauer, et al., The maternal lifestyle study: drug use by meconium toxicology and maternal self-report, Pediatrics 107 (2001) 309–317.
[79] E.J. Cone, M. Hillsgrove, W.D. Darwin, Simultaneous measurement of cocaine, cocaethylene, their metabolites, and "crack" pyrolysis products by gas chromatography-mass spectrometry, Clin. Chem. 40 (1994) 1299–1305.
[80] L.J. Langman, M.W. Bjergum, C.L. Williamson, F.W. Crow, Sensitive method for detection of cocaine and associated analytes by liquid chromatography-tandem mass spectrometry in urine, J. Anal. Toxicol. 33 (2009) 447–455.
[81] K.M. Babu, R.P. Ferm, Hallucinogens, in: G. Flomenbaum, H. Hoffman, N. Lewin (Eds.), Goldfrank's Toxicologic Emergencies, eight ed, McGraw-Hill, New York, 2006, pp. 1202–1211.
[82] G.K. Poch, K.L. Klette, C. Anderson, The quantitation of 2-oxo-3-hydroxy lysergic acid diethylamide (O-H-LSD) in human urine specimens, a metabolite of LSD: comparative analysis using liquid chromatography-selected ion monitoring mass spectrometry and liquid chromatography-ion trap mass spectrometry, J. Anal. Toxicol. 24 (2000) 170–179.
[83] W.H. Porter, Chapter 34: Clinical toxicology, in: C.A. Burtis, E.R. Ashwood, D.E. Bruns (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, fourth ed, Elsevier Saunders, St. Louis, MO, 2006, pp. 1287–1369.
[84] S.A. Reuschel, S.E. Percey, S. Liu, D.M. Eades, R.L. Foltz, Quantitative determination of LSD and a major metabolite, 2-oxo-3-hydroxy-LSD, in human urine by solid-phase extraction and gas chromatography-tandem mass spectrometry, J. Anal. Toxicol. 23 (1999) 306–312.
[85] M. Peat, Workplace drug testing, in: M. Anthony, M.D. Osselton, W. Brian (Eds.), Clarke's Analysis of Drugs and Poisons, vol. 3, Pharmaceutical Press, London, 2004, pp. 68–79.
[86] H.B. Gutstein, H. Akil, Chapter 21: Opioid analgesics, in: L.L. Brunton, J.S. Lazo, K.L. Parker (Eds.), Goodman & Gilman's the Pharmacological Basis of Therapeutics, eleventh ed, McGraw-Hill Companies, Inc, 2006. Available from: http://www.accessmedicine.com/content.aspx?aID = 940653.
[87] J. Lotsch, Opioid metabolites, J. Pain Symptom Manage. 29 (2005) S10–S24.
[88] A.A. Somogyi, D.T. Barratt, J.K. Coller, Pharmacogenetics of opioids, Clin. Pharmacol. Ther. 81 (2007) 429–444.
[89] J. de Leon, L. Dinsmore, P. Wedlund, Adverse drug reactions to oxycodone and hydrocodone in CYP2D6 ultrarapid metabolizers, J. Clin. Psychopharmacol. 23 (2003) 420–421.
[90] T. Tateishi, Y. Krivoruk, Y.F. Ueng, A.J. Wood, F.P. Guengerich, M. Wood, Identification of human liver cytochrome P-450 3A4 as the enzyme responsible for fentanyl and sufentanil N-dealkylation, Anesth. Analg. 82 (1996) 167–172.
[91] C. Pasero, Fentanyl for acute pain management, J. Perianesth Nurs. 20 (2005) 279–284.
[92] S. Thomas, R. Winecker, J.P. Pestaner, Unusual fentanyl patch administration, Am. J. Forensic Med. Pathol. 29 (2008) 162–163.
[93] K.S. Latta, B. Ginsberg, R.L. Barkin, Meperidine: a critical review, Am. J. Ther. 9 (2002) 53–68.
[94] I. Mancini, D.A. Lossignol, J.J. Body, Opioid switch to oral methadone in cancer pain, Curr. Opin. Oncol. 12 (2000) 308–313.
[95] R.B. Raffa, Basic pharmacology relevant to drug abuse assessment: tramadol as example, J. Clin. Pharm. Ther. 33 (2008) 101–108.
[96] S.L. Walsh, K.L. Preston, M.L. Stitzer, E.J. Cone, G.E. Bigelow, Clinical pharmacology of buprenorphine: ceiling effects at high doses, Clin. Pharmacol. Ther. 55 (1994) 569–580.
[97] P.J. Fudala, R.E. Johnson, Development of opioid formulations with limited diversion and abuse potential, Drug. Alcohol. Depend. 83 (Suppl. 1) (2006) S40–S47.
[98] S.C. Armstrong, K.L. Cozza, Pharmacokinetic drug interactions of morphine, codeine, and their derivatives: theory and clinical reality, part I, Psychosomatics 44 (2003) 167–171.

[99] S.C. Armstrong, K.L. Cozza, Pharmacokinetic drug interactions of morphine, codeine, and their derivatives: theory and clinical reality, part II, Psychosomatics 44 (2003) 515–520.
[100] R. Olmedo, Phencyclidine and ketamine, in: L. Goldfrank, N. Flomenbaum, M. Howland, N. Lewin, R. Hoffman, L. Nelson (Eds.), Goldfrank's Toxicologic Emergencies, eighth ed, McGraw Hill, New York, 2006, pp. 1231–1243.
[101] J.T. Cody, S. Valtier, Effects of stealth adulterant on immunoassay testing for drugs of abuse, J. Anal. Toxicol. 25 (2001) 466–470.
[102] J. Kennedy, J.M. Roll, T. Schraudner, S. Murphy, S. McPherson, Prevalence of persistent pain in the U.S. adult population: new data from the 2010 national health interview survey, J. Pain. 15 (2014) 979–984.
[103] L. Manchikanti, B. Fellows, H. Ailinani, V. Pampati, Therapeutic use, abuse, and nonmedical use of opioids: a ten-year perspective, Pain Physician 13 (2010) 401–435.
[104] System N.V.S., Unintended drug overdose by major type of drug, 1999–2007.
[105] P.J. Christo, L. Manchikanti, X. Ruan, M. Bottros, H. Hansen, D. R. Solanki, et al., Urine drug testing in chronic pain, Pain Physician 14 (2011) 123–143.
[106] Control CFD, Prescription drug overdose in the United States: fact sheet, 2015.
[107] S. Atluri, H. Akbik, G. Sudarshan, Prevention of opioid abuse in chronic non-cancer pain: an algorithmic, evidence based approach, Pain Physician 15 (2012) ES177–ES189.
[108] S. Broekmans, F. Dobbels, K. Milisen, B. Morlion, S. Vanderschueren, Pharmacologic pain treatment in a multidisciplinary pain center: do patients adhere to the prescription of the physician? Clin. J. Pain. 26 (2010) 81–86.
[109] J.F. Peppin, S.D. Passik, J.E. Couto, P.G. Fine, P.J. Christo, C. Argoff, et al., Recommendations for urine drug monitoring as a component of opioid therapy in the treatment of chronic pain, Pain Med. 13 (2012) 886–896.
[110] G. Schonwald, What is the role of urine drug testing (UDT) in the management of chronic non-cancer pain with opioids? Pain Med. 13 (2012) 853–856.
[111] G.T. Owen, A.W. Burton, C.M. Schade, S. Passik, Urine drug testing: current recommendations and best practices, Pain Physician 15 (2012) ES119–ES133.
[112] P.J. Jannetto, N.C. Bratanow, W.A. Clark, R.J. Hamill-Ruth, C. A. Hammett-Stabler, M.A. Huestis, et al., Executive summary: American association of clinical chemistry laboratory medicine practice guideline—using clinical laboratory tests to monitor drug therapy in pain management patients, J. Appl. Lab. Med. 2 (2018) 489.
[113] E.M. Ostrea Jr., Understanding drug testing in the neonate and the role of meconium analysis, J. Perinat. Neonatal Nurs. 14 (2001) 61–82. quiz 105-6.
[114] E.M. Ostrea Jr., M.J. Brady, P.M. Parks, D.C. Asensio, A. Naluz, Drug screening of meconium in infants of drug-dependent mothers: an alternative to urine testing, J. Pediatr. 115 (1989) 474–477.
[115] M.A. ElSohly, S. Feng, Delta 9-THC metabolites in meconium: identification of 11-OH-delta 9-THC, 8 beta,11-diOH-delta 9-THC, and 11-nor-delta 9-THC-9-COOH as major metabolites of delta 9-THC, J. Anal. Toxicol. 22 (1998) 329–335.
[116] S.N. Ahanya, J. Lakshmanan, B.L. Morgan, M.G. Ross, Meconium passage in utero: mechanisms, consequences, and management, Obstet. Gynecol. Surv. 60 (2005) 45–56. quiz 73-4.
[117] J. Gareri, J. Klein, G. Koren, Drugs of abuse testing in meconium, Clin. Chim. Acta 366 (2006) 101–111.
[118] D.E. Lewis, C.M. Moore, J.B. Leikin, A. Koller, Meconium analysis for cocaine: a validation study and comparison with paired urine analysis, J. Anal. Toxicol. 19 (1995) 148–150.
[119] J.M. Colby, Comparison of umbilical cord tissue and meconium for the confirmation of in utero drug exposure, Clin. Biochem. 50 (2017) 784–790.
[120] J.M. Colby, B.C. Adams, A. Morad, L.D. Presley, S.W. Patrick, Umbilical cord tissue and meconium may not be equivalent for confirming in utero substance exposure, J. Pediatr. 205 (2019) 277–280.
[121] M.G. Horning, L. Brown, J. Nowlin, K. Lertratanangkoon, P. Kellaway, T.E. Zion, Use of saliva in therapeutic drug monitoring, Clin. Chem. 23 (1977) 157–164.
[122] K. Hold, D. de Boer, J. Zuidema, R.A. Maes, Saliva as an analytical tool in toxicology, Int. J. Drug Test. 1 (1995) 1–36.
[123] I.A. Siegel, The role of saliva in drug monitoring, Ann. N. Y. Acad. Sci. 694 (1993) 86–90.
[124] V. Spiehler, Drugs in saliva, in: A. Moffat, M.D. Osselton, B. Widdop (Eds.), Clarke's Analysis of Drugs and Poisons, third ed, Pharmaceutical Press, London, 2004, pp. 109–123.
[125] F.M. Wylie, H. Torrance, R.A. Anderson, J.S. Oliver, Drugs in oral fluid Part I. Validation of an analytical procedure for licit and illicit drugs in oral fluid, Forensic Sci. Int. 150 (2005) 191–198.
[126] P. Kintz, M. Villain, B. Ludes, Testing for zolpidem in oral fluid by liquid chromatography-tandem mass spectrometry, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 811 (2004) 59–63.
[127] P. Kintz, Value of hair analysis in postmortem toxicology, Forensic Sci. Int. 142 (2004) 127–134.
[128] A.M. Baumgartner, P.F. Jones, W.A. Baumgartner, C.T. Black, Radioimmunoassay of hair for determining opiate-abuse histories, J. Nucl. Med. 20 (1979) 748–752.
[129] R. Kronstrand, S. Forstberg-Peterson, B. Kagedal, J. Ahlner, G. Larson, Codeine concentration in hair after oral administration is dependent on melanin content, Clin. Chem. 45 (1999) 1485–1494.
[130] F. Pragst, M.A. Balikova, State of the art in hair analysis for detection of drug and alcohol abuse, Clin. Chim. Acta 370 (2006) 17–49.
[131] W.A. Baumgartner, V. Hill, Hair analysis for drugs of abuse: decontamination issues, in: I. Sunshine (Ed.), Recent Developments in Therapeutic Drug Monitoring and Clinical Toxicology, Marcel Dekker Inc, New York, 1992, pp. 577–597.
[132] J. Lohrke, T. Frenzel, J. Endrikat, F.C. Alves, T.M. Grist, M. Law, et al., 25 years of contrast-enhanced MRI: developments, current challenges and future perspectives, Adv. Ther. 33 (2016) 1–28.
[133] P. Robert, S. Fingerhut, C. Factor, V. Vives, J. Letien, M. Sperling, et al., One-year retention of gadolinium in the brain: comparison of gadodiamide and gadoterate meglumine in a rodent model, Radiology 288 (2018) 424–433.

[134] F.G. Strathmann, L.M. Blum, Chapter 42: Toxic elements, in: N. Rifai, A.R. Horvath, C.T. Wittwer (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, sixth ed, Elsevier, St Louis, MO, 2018, pp. 888–924.

[135] N.L. Benowitz, Chapter 16: Antidepressants, tricyclic, in: K.R. Olson (Ed.), Poisoning & Drug Overdose, sixth ed, The McGraw-Hill Companies, New York, 2012.

[136] A. Sorisky, D.C. Watson, Positive diphenhydramine interference in the EMIT-st assay for tricyclic antidepressants in serum, Clin. Chem. 32 (1986) 715.

[137] K.W. Ryder, M.R. Glick, The effect of thioridazine on the automatic clinical analyzer serum tricyclic anti-depressant screen, Am. J. Clin. Pathol. 86 (1986) 248–249.

[138] S. Pankey, C. Collins, A. Jaklitsch, A. Izutsu, M. Hu, M. Pirio, et al., Quantitative homogeneous enzyme immunoassays for amitriptyline, nortriptyline, imipramine, and desipramine, Clin. Chem. 32 (1986) 768–772.

[139] F.K. Mohammad, A.S. Alias, O.A. Ahmed, Electrometric measurement of plasma, erythrocyte, and whole blood cholinesterase activities in healthy human volunteers, J. Med. Toxicol. 3 (2007) 25–30.

[140] J.W. Perez, B.G. Pantazides, C.M. Watson, J.D. Thomas, T.A. Blake, R.C. Johnson, Enhanced stability of blood matrices using a dried sample spot assay to measure human butyrylcholinesterase activity and nerve agent adducts, Anal. Chem. (2015).

Self-assessment questions

1. A previously healthy 4-year-old child was brought to the local hospital because of nausea, vomiting, and increased breathing difficulty. An empty container of roach poison was found near the child. The child had tearing bilaterally, rhinorrhea, gurgling sonorous respirations when supine, and was incontinent (urine/stool). Based on her clinical presentation, the most likely explanation for the toxicity and treatment or antidote would be:
   a. carbamate poisoning; fresh frozen plasma
   b. arsenic poisoning; ascorbic acid
   c. organophosphate poisoning; atropine/pralidoxime
   d. chromium poisoning; EDTA chelation therapy
   e. iron poisoning; deferoxamine
2. A 78-year-old man was found unresponsive in his bed after the first cold day of winter. There is a kerosene heater still on from the previous night. On attempts to arouse him, officers note the cherry hue of his lips, cheek, and mucous membranes. The most likely mechanism of his death was:
   a. accidental ingestion of ethylene glycol
   b. hepatic necrosis with fatty changes
   c. inhibition of fatty acid synthesis
   d. binding of carbon monoxide to hemoglobin
   e. inhibition of iron incorporation into hemoglobin
3. Which of the following compounds may be responsible for a combined anion and osmol gap?
   a. aspirin
   b. cyanide
   c. iron
   d. isoniazid
   e. ethylene glycol
4. A patient is brought into the ED with the following symptoms: coma, respiratory depression, and pinpoint pupils. He is given naloxone and responds. The expected result on the urine toxicology screen is that it will be positive for:
   a. PCP
   b. cocaine
   c. amphetamine
   d. opiates
   e. marijuana
5. A 6-year-old male is brought to the Emergency Room by his mother. The child has been nauseous and vomiting for 2 days and complains his stomach hurts. His physical exam showed the following: $T = 98.6°F$, blood pressure = 110/70 mm Hg, heart rate = 103 bpm, and respiratory rate = 22 breaths/min. He was jaundiced and diaphoretic. Laboratory testing showed the following:

		Reference range
• Sodium	130 mmol/L	(135–145)
• Potassium	4.0 mmol/L	(3.4–4.7)
• Chloride	101 mmol/L	(98–109)
• Glucose	88 mg/dL	(75–99)
• ALT	1200	(<45)
• AST	1050	(10–45)
• PT-INR	3.1	(0.8–1.2)
• Urine drug of abuse panel	Negative	(Negative)
• Salicylate	Negative	(Negative)
• Acetaminophen	35 mg/L	(10–30)

The most likely diagnosis/treatment for the child is:

a. salicylate overdose/activated charcoal
b. acetaminophen overdose/*N*-acetylcysteine (Mucomyst)
c. alcohol overdose/dialysis
d. digoxin overdose/Digibind
e. warfarin overdose/vitamin K

Answers

1. c
2. d
3. e
4. d
5. b

Chapter 53

Pharmacogenomics

Gwendolyn A. McMillin

Department of Pathology at the University of Utah School of Medicine, ARUP Laboratories; Salt Lake City, UT, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Explain how pharmacogenetics can help guide drug and dose selection decisions.
- Describe examples of gene–drug pairs used to qualify a patient for drug therapy.
- Describe examples of gene–drug pairs used to optimize drug dose.
- Discuss limitations of pharmacogenetic testing.

Introduction

Pharmacology is the study of how drugs exert both desirable (e.g., therapeutic) and undesirable (e.g., toxic) effects. The two major disciplines that explain the processes associated with drug response are pharmacokinetics, describing what the body does to a drug, and pharmacodynamics, describing what the drug does to the body. Proteins commonly involved in pharmacokinetics include drug metabolizing enzymes and drug transporters. Proteins commonly involved in pharmacodynamics include signaling pathways (e.g., enzymes, receptors, and ion channels) and immune mediators. Regardless of the downstream effect, each protein is derived from genes that may exhibit variation, and that variation tends to cluster in certain populations. Common variations include nucleotide base changes, insertions or deletions of nucleotides, and copy number variations (CNVs). Such a genetic variation may influence protein expression and/or function. The association of specific gene variants with discrete aspects of pharmacokinetics and/or pharmacodynamics is known as pharmacogenetics. The genes involved in the pharmacogenetic associations are called pharmacogenes. Many drug–gene pairs are now described that can be used to guide drug and dose selection to maximize therapeutic response and minimize risk of toxicity for an individual patient. The recognition that many genes within a genome can be involved in the response to a drug has contributed to the vision of pharmacogenomics, wherein several pharmacologically relevant variants are detected and characterized for many genes, and eventually the whole genome. As variation in drug response is rarely explained by only one gene, and as many individuals take more than one drug, pharmacogenomics has great potential to improve patient therapy decisions over what a single drug–gene pair can accomplish. That said, the terms pharmacogenetics and pharmacogenomics are often used interchangeably.

Drug response

A person's unique response to a drug is the phenotype that results from the coordination of pharmacokinetic and pharmacodynamic processes for that drug or combination of drugs, in that person. Possible drug response phenotypes include the following: (1) efficacy without toxicity; (2) efficacy with toxicity; (3) toxicity without efficacy; or (4) neither efficacy nor toxicity. Responses may be dose-dependent or may occur independent of dose. Recognition of drug response phenotypes that cluster within families has led to characterization of many pharmacogenetic associations. Based on these associations, genetic tests have been developed to predict whether a person is likely to respond to a drug, whether a person is at increased risk for an adverse event, and/or whether a person might respond best to a nonstandard dose of a particular drug, before the drug is ever administered. However, drug response phenotypes are not typically explained by genetics alone.

The drug response phenotype is also influenced by nongenetic factors, such as age, sex, clinical health (e.g., kidney function and liver function), comedications, diet, and lifestyle factors (e.g., use of alcohol and/or nicotine). Specific drug formulation and route of administration may also influence response to drugs. Thus pharmacogenetics is but one tool that is used to inform and guide drug

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00053-3

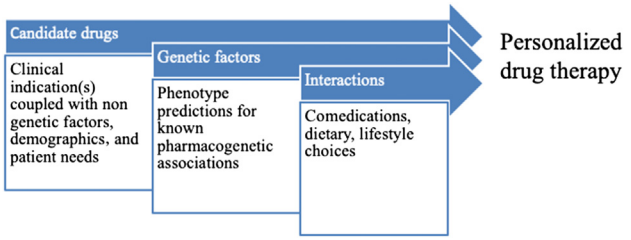


FIGURE 53.1 Many factors are used to select a drug and dose of drug for an individual patient. This schematic illustrates that candidate drugs for a clinical indication/diagnosis are narrowed based on genetic and nongenetic factors, including the potential for drug–drug and food–drug interactions.

and dosing decisions. As illustrated in Fig. 53.1, there may be several candidate drugs available to treat a patient for a particular clinical indication based on diagnosis and nongenetic factors. By incorporating genetic predictions of drug response along with clinical factors, patient demographics and the potential for drug–drug and drug–food interactions, therapy can be personalized, making both drug and dosing decisions more accurate.

Pharmacogenetic testing

Clinical pharmacogenetic testing is currently focused on detection of discrete genetic variants that are associated with well-defined aspects of pharmacokinetics and pharmacodynamics for a specific drug or class of drugs. As illustrated in Fig. 53.2, a laboratory test is designed to detect specific genetic variants (e.g., single nucleotide variants, insertions/deletions, and CNVs), referred to as genotypes. This information is used to assign alleles, often making assumptions about haplotypes/diplotypes (i.e., whether multiple variants occur on just one chromosome or whether such variants are divided between two chromosomes). The reason that haplotypes are often inferred is that most analytical methods can detect a gene variant but the method cannot determine the phase (chromosome position) of each variant. As such, specialized testing that can discriminate phase and frequency statistics is often used to predict the specific chromosome on which the pattern of genetic variants is likely to reside. The alleles are then assembled as a diplotype; this information is then used to predict the phenotype. Translation tables for both alleles and phenotypes are available through many resources, such as the Pharmacogenomics Knowledgebase (PharmGKB) [1]. Standardization of terms to define the predicted phenotype has also been proposed [2].

Some pharmacogenes are relatively simple to detect. For example, the association of many human leukocyte antigen (HLA) gene variants with potentially life-threatening drug hypersensitivity reactions requires a test that can identify the presence of at least one variant allele. The mode of inheritance is autosomal dominant for the well-known association between *HLA-B*57:01* and risk of hypersensitivity to abacavir; so, it does not matter whether one or two affected alleles are detected. The potentially life-threatening immune-mediated reaction that can precipitate when a carrier of the *HLA-B*57:01* allele is challenged with abacavir occurs regardless of dose or variation in pharmacokinetics, making implementation a rather straightforward process. Thus a carrier that is naïve to abacavir is disqualified from initiating that therapy. Alternative drugs are offered instead [3].

The *SLCO1B1* gene is also relatively simple from an analytical perspective, in that a test designed to detect only one variant is sufficient to predict increased risk of myopathy from simvastatin therapy, a commonly prescribed drug used for cholesterol reduction. The organic anion transporter polypeptide 1B1, encoded by *SLCO1B1*, transports active simvastatin acid from the blood stream into the liver. Detection of the variant (rs4149056, c.521 T > C) is associated with reduced transporter function, resulting in an accumulation of the drug and subsequent muscle toxicity. One copy of the variant allele predicts decreased transporter function; two copies of the variant allele predict poor transporter function. This is in contrast to the *HLA-B*57:01* example above, because the mode of inheritance of *SLCO1B1* and most other pharmacogenes are autosomal codominant, leading to a range of phenotype predictions [2]. A carrier of a variant *SLCO1B1* allele could minimize risk of the toxicity by using a lower dose or an alternate drug. Monitoring creatine kinase concentrations may also be useful for detecting and managing muscle toxicity associated with statin therapy [4].

Common nomenclature systems in pharmacogenetics classify combinations of sequence variants as star (*) alleles, where *1 usually reflects the reference allele (commonly referred to as "normal"), and numbered star alleles are assigned sequentially as new variants are identified [5]. Variants are defined, and nomenclature is archived by repositories, such as the Pharmacogene Variation (PharmVar) Consortium. Few clinical laboratories offer complete gene sequencing, phase, and CNV determinations for pharmacogenes. Instead, most testing is targeted, designed to detect only select variants of known clinical significance. Pragmatically, the *1 allele is assigned by default when none of the targeted variants are detected. Therefore the true accuracy of a *1 allele designation is unlikely and depends on whether the assay detects all possible variants that an individual may carry. The star (*) approach to nomenclature has been adopted for many pharmacogenes. A summary of recommendations for nomenclature reporting pharmacogene data is shown in Table 53.1 [6].

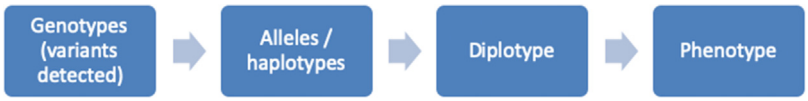


FIGURE 53.2 Translation of pharmacogenetic results relies on comparison of results with consensus nomenclature to define alleles, haplotypes, diplotypes, and, ultimately, the predicted phenotype. The phenotype may also be influenced by nongenetic factors, such as drug−drug interactions.

TABLE 53.1 General recommendations for pharmacogenetics nomenclature standardization and reporting [6].

Category	Recommendation
Naming the pharmacogenetic variants	Use HUGO Gene Nomenclature Committee nomenclature to specify the gene
	Report variants using Human Genome Variation Society nomenclature
	Use a Locus Reference Genomic reference sequence gene and/or a specific Human Genome Reference Assembly as a reference sequence. Both the reference sequence number and version number of the sequence (if applicable) should be indicated
	Report reference sequence numbers from the National Center for Biotechnology Information dbSNP database, the PharmVar database, or other reliable source
	Use haplotype translation tables to convert star (*) alleles or other legacy nomenclature to fully specified Human Genome Variation Society nomenclature for each variant in the haplotype. Note that nomenclature can be subject to change based on new findings and consensus
Test reporting	Report each variant and/or haplotype observed
	List variants and haplotypes that can be detected by the test (specific sites for genotyping tests or regions for sequencing-based tests)
	Describe the test, including limitations such as types of variants that cannot be detected
	The test description should be made publicly available on the laboratories' website, and/or by registration of the test in the NIH Genetic Testing Registry

Genes that code for drug metabolizing enzymes are among the best characterized and most commonly cited pharmacogenes. An example of nomenclature used to classify cytochrome P450 (CYP) drug metabolizing enzymes and the associated genes is shown in Fig. 53.3. *CYP2D6*, is one of the most clinically relevant pharmacogenes, but is also one of the most complicated. For example, the *CYP2D6*4* allele is defined primarily by the c.1847 G > A (rs3892097) variant, and subtypes are determined by combinations of additional variants [7]. The *4A subtype (PharmVar *CYP2D6**4.001) is defined by a combination of seven variants. There are over 100 unique star (*) alleles and many more subtypes defined for *CYP2D6*, making this an extremely difficult gene to interrogate with accuracy and specificity. That said, CYP subtypes do not usually affect the phenotype prediction and are most important for characterizing inheritance patterns in families, or for supporting research studies.

Classification of alleles and associated haplotypes/diplotypes is used for predicting the phenotype. When considering genes that code for drug metabolizing enzymes, each allele is classified based on predicted

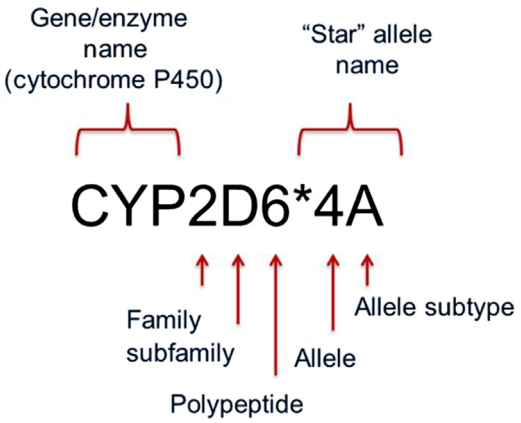


FIGURE 53.3 Description of how the cytochrome P450 drug metabolizing enzymes and associated genes are named using the isozyme subtype 4A of *CYP2D6* as an example.

enzyme function, such as normal, decreased function, no function, or increased function. A diplotype, which may be presumed based on whether phase analysis to delineate haplotypes is performed, is further translated into predicted drug metabolizer status (i.e., phenotype), which

TABLE 53.2 Examples of standardized terms for pharmacogenetic phenotype predictions from diplotypes published by the Clinical Pharmacogenetics Implementation Consortium [2].

Category	Phenotype term for diplotype	Functional definition	Genetic definition	Example gene diplotype
Drug metabolizing enzymes	Ultrarapid metabolizer	Enzyme activity higher than rapid metabolizers	More than two functional alleles or two increased function alleles	*CYP2C19*17/*17*
	Rapid metabolizer	Increased enzyme activity	Combinations of normal and increased function alleles	*CYP2C19*1/*17*
	Normal metabolizer	Fully functional enzyme activity	Two functional alleles or combinations of functional and decreased function alleles	*CYP2C19*1/*1*
	Intermediate metabolizer	Reduced enzyme activity	Combinations of normal function, decreased function, and/or no function alleles	*CYP2C19*1/*2*
	Poor metabolizer	Little-to-no enzyme activity	Combinations of no function alleles and/or decreased function alleles	*CYP2C19*2/*2*
Transporters	Increased function	Increased transporter function	One or more increased function alleles	*SLCO1B1*1/*14*
	Normal function	Fully functional transporter	Two functional alleles or combinations of functional and decreased function alleles	*SLCO1B1*1/*1*
	Decreased function	Reduced transporter function	Combinations of normal function, decreased function, and/or no function alleles	*SLCO1B1*1/*5*
	Poor function	Little-to-no transporter function	Combinations of no function alleles and/or decreased function alleles	*SLCO1B1*5/*5*

could be normal, intermediate, poor, rapid, or ultrarapid. For example, a poor metabolizer is predicted when two copies of nonfunctional alleles are inherited (e.g., *CYP2D6*4/*4*). Terms for phenotype predictions are now standardized for some pharmacogenes, as shown in Table 53.2 [2].

The *4 is the most common no-function *CYP2D6* allele. However, one of the variants used to define all, but one currently recognized *4 allele subtype (c.100 C > T, rs1065852), is also used to define dozens of other *CYP2D6* alleles, and some of which exhibit different phenotypic characteristics than the *4 allele (e.g., *CYP2D6*10*). This is an example wherein a genotype may be misclassified as the wrong allele and subsequent diplotype, such as assigning *4/*10 instead of *1/*4. Misclassification of alleles may occur with targeted testing and when consensus nomenclature is not consulted. Such misclassification has potential clinical significance if it affects the phenotype prediction. In this case, a *4/*10 would be classified as an intermediate metabolizer, whereas a *1/*4 would be classified as a normal metabolizer. Another analytical challenge is that many newly recognized *CYP2D6* alleles are gene hybrids that incorporate pieces of pseudogenes, such as *CYP2D7*, into a portion of the *CYP2D6* gene (e.g., *CYP2D6*36*). The *CYP2D6*36* is relatively common in Asian populations, but many gene hybrids are rare and are still being recognized, defined, and characterized.

Many important variants reside in noncoding regions or occur as a consequence of CNVs [1]. Few traditional or massively parallel sequencing methods are optimized to detect all clinically relevant pharmacogenetic information. As a consequence of these analytical challenges, there is an overall lack of standardization in the specific alleles detected among laboratories for *CYP2D6*, and for some other complicated pharmacogenes. Complementary testing that provides more comprehensive detection of genetic variants (e.g., full gene sequencing versus targeting only common variants) or testing that characterizes the phenotype directly (e.g., therapeutic drug monitoring, metabolic ratios, and enzyme activity testing) can refine the clinical value of pharmacogenetic testing, particularly for complicated pharmacogenes, such as *CYP2D6*.

TABLE 53.3 Tiers of testing proposed for *CYP2C19* by the Association for Molecular Pathology [8].

Tier	Star allele	Allele function	Defining variant	Common HGVS nomenclature (NM_00769.2)
1	*2	No function	rs4244285	c.681 G > A
	*3	No function	rs4986893	c.636 G > A
	*17	Increased function	rs12248560	c. − 806 C > T
2	*4A	No function	rs28399504	c.1 A > G
	*4B	No function	rs12248560, rs28399504	c. − 806 C > T, c.1 A > G
	*5	No function	rs56337013	c.1297 C > T
	*6	No function	rs72552267	c.395 G > A
	*7	No function	rs72558186	c.819 + 2 T > A
	*8	No function	rs41291556	c.358 T > C
	*9	Decreased function	rs17884712	c.431 G > A
	*10	Decreased function	rs6413438	c.680 C > T
	*35	No function	rs12769205	c.332−23 A > G

HGVS; Human Genome Variation Society.

Efforts to harmonize pharmacogenetic testing are underway, such as the Association for Molecular Pathology's recommendations for clinical *CYP2C19* testing. Well-characterized alleles were considered based on availability of accurate testing, reference materials, multiethnic allele frequencies, and clinical guidance documents for interpretation and implementation. As shown in Table 53.3, there are three alleles that were recommended for "Tier 1" testing. Tier 1 testing defined the minimum content that a clinical laboratory should offer, and could offer, based on meeting the criteria described above. "Tier 2" testing expands on that list to include variant alleles of clinical importance to a smaller group of individuals, or for which analytical characteristics are not yet mature [8]. Consistent with the recommendations in Table 53.1, the variant alleles listed in Table 53.3 are defined based on primary causative variant [reference sequence (rs) number and common designation, with rf], along with the predicted function of each allele. Clinical laboratories that offer pharmacogenetic testing can be found through the voluntary National Institutes of Health Genetic Testing Registry [9].

In addition to concerns about accuracy of haplotype and diplotype assignments used to predict phenotype that were raised above, the actual drug response phenotype is affected by other genetic factors not detected by the test (e.g., other pharmacogenes involved in drug response), as well as demographic factors (e.g., age and sex), clinical factors (e.g., body size and kidney function), and nongenetic factors, such as drug−drug and food−drug interactions. For example, a person who inherits no *CYP2D6* variants would be anticipated to exhibit normal metabolism and achieve therapeutic concentrations of CYP2D6 drug substrates, but a drug may not be effective if the physiology to support the drug's mechanism of action is impaired or absent (e.g., opioid receptors that cannot respond to a opioid substrate that requires metabolic activation via CYP2D6). Also, a person who inherits no *CYP2D6* variants could exhibit a poor metabolizer phenotype if comedicated with an inhibitor of CYP2D6, such as fluoxetine. The US Food and Drug Administration (FDA) classifies drugs as strong, moderate, or weak inhibitors [10]. These coadministered drugs can elicit a poor metabolizer phenotype by inhibiting CYP2D6. A genetically predicted intermediate metabolizer is particularly vulnerable to drug−drug interactions. In addition, some supplements and food can inhibit CYP enzymes. For example, grapefruit juice is recognized to inhibit CYP3A4 activity, leading to reduced metabolism of associated drug substrates. As such, assignment of drug metabolizing enzyme pharmacogenetics to phenotype predictions and associated clinical implementation is multifactorial.

Pharmacogenetics implementation

The Clinical Pharmacogenetics Implementation Consortium (CPIC) and other organizations, such as the Dutch Pharmacogenomics Working Group (DPWG), publish guidelines for translation of the pharmacogenetic

TABLE 53.4 Examples of pharmacogenetic associations for which dosing guidelines were published by the Clinical Pharmacogenetics Implementation Consortium [1].

Drug	Gene	Utility		FDA label classification
		Qualify	Dose selection	
Carbamazepine	*HLA-B*15:02*	X		*Requirement* for genetic testing
Ivacaftor	*CFTR*	X		*Requirement* for genetic testing
Rasburicase	*G6PD*	X		*Requirement* for genetic testing
Abacavir	*HLA-B*57:01*	X		*Recommendation* for genetic testing
Clopidogrel	*CYP2C19*	X		*Recommendation* for genetic testing
Thiopurines	*TPMT*	X	X	*Recommendation* for genetic testing; includes azathioprine, mercaptopurine, and thioguanine
Fluoropyrimidines	*DPYD*	X	X	*Actionable*; includes 5-fluorouracil, capecitabine, and tegafur
Tricyclic antidepressants	*CYP2D6* and *CYP2C19*		X	*Actionable*; includes amitriptyline, doxepin, desipramine, imipramine, nortriptyline, and trimipramine; value of *CYP2D6* and/or *CYP2C19* is drug specific
Selective serotonin reuptake inhibitors	*CYP2D6* and *CYP2C19*		X	*Actionable*; includes citalopram, escitalopram, fluvoxamine, and paroxetine; value of *CYP2D6* and/or *CYP2C19* is drug specific
Opioids	*CYP2D6*	X	X	*Actionable;* includes codeine and tramadol
Phenytoin	*HLA-B*15:02* and *CYP2C9*	X	X	*Actionable*
Simvastatin	*SLCO1B1*	X	X	*Informative*
Tacrolimus	*CYP3A5*		X	Not currently addressed
Warfarin	*CYP2C9*		X	*Actionable*; includes dosing information
	VKORC1			
	CYP4F2			

TPMT, Thiopurine s-methyltransferase.

information to phenotype predictions and, subsequently, recommendations for specific drug, dosing, and management decisions unique to an individual patient. These guidelines do not make recommendations about when or whether testing should occur, but rather focus on evidence-based considerations that may better inform medication decisions when pharmacogenetic information is available. Examples of gene–drug pairs for which dosing guidelines are published by the CPIC are shown in Table 53.4, along with comments regarding the classification of the association by the FDA. The FDA has adopted language in drug labeling, indicating that testing is "required," "recommended," "actionable," or "informative."

Gene-based dosing guidelines typically represent gene–drug pairs that are associated with the highest level of evidence for clinical use. The CPIC levels of evidence (A–D) are assigned based on endorsement for a specific gene–drug association by professional medical associations and/or the PharmGKB assignments of evidence at Level 1 or 2. The PharmGKB grades gene–drug associations on a scale from 1 (highest level of evidence) to 4 (lowest level of evidence). Level-1 associations are replicated in multiple clinical cohorts with statistically significant differences in outcomes when the pharmacogene is used to inform medication decisions. The CPIC assigns gene–drug associations as Level "A" when the evidence supports a prescribing action, including alternate therapy or changes in dosing. A CPIC grade of "B" indicates the evidence is "moderate," but that prescribing actions may be appropriate. The CPIC also grades gene–drug associations as Level C or D when the PharmGKB classifies a

TABLE 53.5 Examples of pharmacogenetic-guided dosing recommendations published by the Clinical Pharmacogenetics Implementation Consortium [1].

Example result	Phenotype prediction	Implications for outcome	Dosing recommendations
HLA-B*57:01			
Negative	Not a carrier	"Normal" risk of abacavir hypersensitivity	Use label-recommended dosage and administration
Positive	Carrier	Increased risk of abacavir hypersensitivity	Select alternate drug; avoid use of abacavir
TPMT			
*1/*1	Normal metabolizer	"Normal" (low) concentrations of thioguanine nucleotide metabolites	Use label-recommended dosage and administration. Adjust based on degree of myelosuppression and disease-specific guidelines. Allow 2–4 weeks to reach steady state after each dose adjustment
		Note that thioguanine nucleotide metabolite concentrations with thioguanine are 5–10× higher than with mercaptopurine or azathioprine	
*1/*2, *1/*3A, *1/*3B, *1/*3C, and *1/*4	Intermediate metabolizer	Moderate-to-high concentrations of thioguanine nucleotide metabolites	Reduce dose by 30%–50% and adjust based on degree of myelosuppression and disease-specific guidelines. Allow 2–4 weeks to reach steady state after each dose adjustment
*3A/*3A, *2/*3A, *3C/*3A, *3C/*4, *3C/*2, and *3A/*4	Poor metabolizer	Extremely high concentrations of thioguanine nucleotide metabolites	Reduce daily dose by 10-fold and dose thrice weekly instead of daily
		Fatal toxicity possible without dose decrease	Adjust based on degree of myelosuppression and disease-specific guidelines. Allow 4–6 weeks to reach steady state after each dose adjustment. For nonmalignant conditions, consider nonthiopurine alternative

TPMT, Thiopurine s-methyltransferase.

gene as Level 1 or 2 but no prescribing actions are recommended based on genetics, and/or when alternative medications are unclear. The discriminating factors between Levels C and D relate to whether testing is common (C) or rare (D) and whether the pharmacogene is associated with CPIC Level A or B for other drugs: yes for C, no for D. The PharmGKB Level-3 gene–drug associations are supported by a single statistically significant study (not replicated yet), and Level-4 gene–drug associations are supported by only case reports, nonsignificant studies, or in vitro evidence only. These evidence levels and associated dosing guidelines are updated as new data become available.

Examples of specific CPIC dosing recommendations are shown in Table 53.5. The CPIC, DPWG, FDA drug labeling, and other sources of pharmacogenetic information are summarized and updated by the PharmGKB [1]. The FDA has included pharmacogenetic information in labels for more than 300 drugs. The European Medicines Agency, Health Canada (Santé Canada), and the Pharmaceuticals and Medical Devices Agency of Japan have also included pharmacogenetic information in labeling for hundreds of drugs. The PharmGKB and other public resources offer many educational resources regarding clinical associations with pharmacogenetic variants, tools for allele and phenotype assignments, allele frequency tables, pathways, and tools for implementation of pharmacogenetics in a clinical setting. Performing pharmacogenetic testing for well-characterized gene–drug pairs, such as those in Table 53.4, can improve accuracy of many pharmacotherapy decisions before a drug is ever administered. However, dose optimization will still rely largely on trial and error, results of related laboratory and clinical monitoring, supplementary laboratory biomarkers, patient compliance, and patient feedback.

Guidance for clinical implementation of pharmacogenetic associations today falls into two major categories: qualification for therapy or dose optimization. The

primary utility is indicated in Table 53.4 for select gene−drug pairs, to serve as examples of how pharmacogenetic information may be used. First, a pharmacogenetic test may qualify a person for drug therapy by either predicting risk of adverse drug reactions or response/nonresponse. As discussed above, detection of at least one *HLA-B*57:01* allele would disqualify a patient for abacavir therapy due to increased risk of a potentially life-threatening hypersensitivity reaction. See Table 53.5 for examples of guidance comments from the CPIC for *HLA-B*57:01* test results. Similarly, detection of an *HLA-B*15:02* allele would disqualify a patient for therapy with carbamazepine or phenytoin (see Table 53.4). These hypersensitivity reactions are not dose-dependent, so cannot be avoided by adjusting dose for a vulnerable patient. *CYP2C19* testing is another example of a gene wherein the presence of a loss of function variant may disqualify a patient from treatment with drugs, such as clopidogrel, which requires metabolic activation, mediated primarily by CYP2C19, to be effective. As a result, patients who are CYP2C19 poor metabolizers are unlikely to respond desirably to clopidogrel [11]. In contrast, testing *CFTR* is used to qualify patients for ivacaftor, because this drug is the first FDA-approved therapeutic developed to target a specific *CFTR* defect. Ivacaftor works by improving the transport of sodium through the CFTR ion channel, which is only effective in patients that have impaired CFTR conductance. The specific variants and combinations of variants associated with response are included in the CPIC guideline [12].

Pharmacogenetic testing results can also be used to adjust the initial dose. Most examples of dose optimization are based on drug metabolism. For extremes of metabolic phenotypes, such as ultrarapid and poor metabolizers, dosing recommendations frequently suggest avoiding drugs that require the affected enzyme. A drug that is inactivated by metabolism may accumulate to potentially toxic blood concentrations in a poor metabolizer. An ultrarapid metabolizer may inactivate the drug to an extent that blood concentrations never become therapeutic. An example of this scenario is provided by CYP2D6 and nortriptyline, wherein drug avoidance is recommended for both poor and ultrarapid metabolizers [13]. In contrast, a drug that is activated by metabolism may produce far more active drug than intended in an ultrarapid metabolizer. For example, codeine is converted to the much more potent drug morphine, primarily based on a reaction mediated by CYP2D6. As such, an ultrarapid metabolizer is at risk of an unintentional overdose after receiving a standard dose of codeine. For a poor metabolizer, codeine may never be activated sufficiently for any response. The CPIC guidelines for codeine suggest alternative drugs for both ultrarapid and poor CYP2D6 metabolizers [14]. For intermediate metabolizer phenotypes, a dose adjustment may compensate for the anticipated change in metabolic activation or inactivation, but the impact of the impaired metabolism is drug-dependent. For nortriptyline, the CPIC recommends a 25% dose reduction for CYP2D6 intermediate metabolizers. The dose can then be optimized with therapeutic drug monitoring. However, no specific dose adjustment is recommended for codeine in CYP2D6 intermediate metabolizers.

Another classic example that falls into this second category of clinical implementation is the thiopurine s-methyltransferase (TPMT) enzyme, coded by a gene of the same name, and thiopurine toxicity. With normal TPMT activity, approximately 90% of thiopurine drugs (e.g., azathioprine, mercaptopurine, and thioguanine) are inactivated. As such, a TPMT poor metabolizer does not inactivate thiopurine drugs and requires a reduction in dose of approximately 90%. The recommendation for TPMT intermediate metabolizer is to reduce the standard dose by 30%−50%. See Table 53.5 for examples of guidance comments from the CPIC for *TPMT* testing to adjust the dose of thiopurine drugs. In this case, both phenotype (e.g., enzyme activity and metabolic ratios of the drugs) and genotype testing can help to inform and optimize dosing. The most recent revision of the CPIC guideline for thiopurine drugs also includes *NUDT15*, which is a gene that codes for nudix (nucleoside diphosphate-linked moiety X)-type motif 15 [15]. Like *TPMT*, nudix is involved in detoxification of active thiopurine metabolites, and loss-of-function variants in *NUDT15* are also associated with an increased risk of dose-related toxicity. The *TPMT* and *NUDT15* variants increase the risk of dose-related toxicity independently, and will exert additive effects. As such, variant alleles of either one or both genes are associated with recommendations for dose reduction. Optimization of dose may require a combination of clinical monitoring and therapeutic drug monitoring.

Limitations

Limitations of pharmacogenetics and pharmacogenomics testing reflect the evolution of this field. The analytical and/or translational concerns addressed throughout this chapter should guide evaluation of testing and interpretive tools in the clinical marketplace. Resources such as CPIC guidelines and FDA-approved labeling recommendations have been incorporated into clinical decision support tools, including alert messages in the electronic medical record and dose calculators. In addition, many laboratories have adopted proprietary algorithms that may provide recommendations based on single genes or combinations of genes. The latter approach is termed “combinatorial” and is currently most prevalent for informing psychotropic medication decisions. Some of the combinatorial

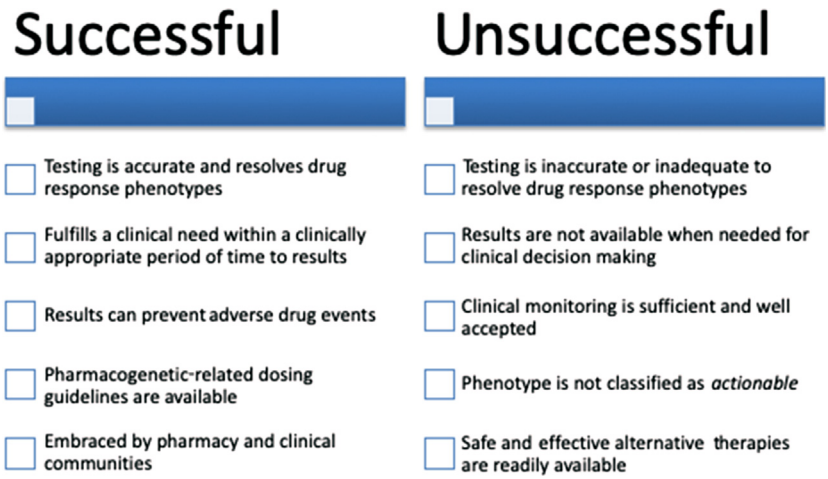


FIGURE 53.4 Examples of factors affecting the successes and limitations of pharmacogenetic applications.

algorithms have been evaluated in randomized clinical trials but most have not. Some algorithms also incorporate clinical, lifestyle, pharmacy admissions, and demographic information to predict risk of adverse drug events or therapeutic failure, particularly for patients receiving polypharmacy. Traditional laboratory testing, including pharmacogenetics testing, requires involvement of the provider to order and interpret a test.

Recently, direct to consumer testing (DTC) has become available for pharmacogenetics. Direct patient access to pharmacogenetic information improves portability of results and may facilitate engagement of multiple medical specialties. DTC testing also helps empower the consumer with information that could impact medication decisions, including selection of nonprescribed medications and supplements. This empowerment could be beneficial to the consumer, but could also introduce risk to the consumer if he or she chooses to change medication and/or dosing decisions independent of clinical providers. Regardless of how results are delivered, pharmacogenetic testing results need to improve patient care beyond what can be achieved with existing standards of care to be successful, and need to engage a multidisciplinary team, representing all stakeholders for a specific application. Examples of these and other factors that influence success of pharmacogenetic testing applications are summarized in Fig. 53.4.

Clinical pharmacogenetic laboratories often provide an interpreted phenotype based on the genotype results. Laboratories that are accredited by the College of American Pathologists (CAPs) are required to include a summary of methods and the variants that can be detected on the report, but, overall, the lack of consensus for pharmacogenetic nomenclature (e.g., *alleles, rs#, Human Genome Variation Society, diplotype versus haplotype, etc.), and variable assay designs add to the complexity of analyzing and reporting results from pharmacogenetic assays. Proficiency testing offered through the CAPs has demonstrated both successes and failures in performance among laboratories performing pharmacogenetic testing [16]. The recommendations indicated in Table 53.1 should be considered when developing a laboratory report for pharmacogenetic testing results.

References

[1] M. Whirl-Carrillo, E.M. McDonagh, J.M. Hebert, L. Gong, K. Sangkuhl, C.F. Thorn, et al., Pharmacogenomics knowledge for personalized medicine, Clin. Pharmacol. Ther. 92 (4) (2012) 414–417.

[2] K.E. Caudle, H.M. Dunnenberger, R.R. Freimuth, J.F. Peterson, J. D. Burlison, M. Whirl-Carrillo, et al., Standardizing terms for clinical pharmacogenetic test results: consensus terms from the Clinical Pharmacogenetics Implementation Consortium (CPIC), Genet. Med. 19 (2) (2017) 215–223.

[3] W.H. Chung, S.I. Hung, Y.T. Chen, Human leukocyte antigens and drug hypersensitivity, Curr. Opin. Allergy Clin. Immunol. 7 (2007) 317–323.

[4] R.A. Wilke, L.B. Ramsy, S.G. Johnson, W.D. Maxwell, H.L. McLeod, D. Voora, et al., The Clinical Pharmacogenetics Implementation Consortium: CPIC guideline for SLCO1B1 and simvastatin-induced myopathy, Clin. Pharmacol. Ther. 92 (1) (2012) 112–117.

[5] J.D. Robarge, L. Li, Z. Desta, A. Nguyen, D.A. Flockhart, The star-allele nomenclature: retooling for translational genomics, Clin. Pharmacol. Ther. 82 (2007) 244–248.

[6] L.V. Kalman, J.A. Agúndez, M.L. Appell, J.L. Black, G.C. Bell, S. Boukouvala, et al., Pharmacogenetic allele nomenclature: international workgroup recommendations for test result reporting, Clin. Pharmacol. Ther. 99 (2) (2016) 172–185.

[7] S.C. Sim, M. Ingelman-Sundberg, The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects, Hum. Genomics 4 (2010) 278–281.

[8] V.M. Pratt, L.D. Tredici, H. Hachad, Y. Ji, L.V. Kalman, S.A. Scott, et al., Recommendations for clinical CYP2C19 genotyping

allele selection: a report of the Association for Molecular Pathology, J. Mol. Diagn. 20 (3) (2018) 269–276.

[9] W.S. Rubinstein, D.R. Maglott, J.M. Lee, B.L. Kattman, A.J. Malheiro, M. Ovetsky, et al., The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency, Nucleic Acids Res. 41 (2013) D925–D935.

[10] U.S. Food and Drug Administration, Drug development and drug interactions: table of substrates, inhibitors and inducers. <http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DevelopmentResources/DrugInteractionsLabeling/ucm093664.htm>, (Accessed December, 2018).

[11] S.A. Scott, K. Sangkuhl, C.M. Stein, J.S. Hulot, J.L. Mega, D. M. Roden, et al., Clinical Pharmacogenetics Implementation Consortium guidelines for CYP2C19 genotype and clopidogrel therapy: 2013 update, Clin. Pharmacol. Ther. 94 (2013) 317–323.

[12] J.P. Clancy, S.G. Johnson, S.W. Yee, E.M. McDonagh, K.E. Caudle, T.E. Klein, et al., Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for ivacaftor therapy in the context of CFTR genotype, Clin. Pharmacol. Ther. 95 (2014) 592–597.

[13] J.K. Hicks, J.J. Swen, C.F. Thorn, K. Sangkuhl, E.D. Kharasch, V. L. Ellingrod, et al., Clinical Pharmacogenetics Implementation Consortium guideline for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants, Clin. Pharmacol. Ther. 93 (2013) 402–408.

[14] K.R. Crews, A. Gaedigk, H.M. Dunnenberger, J.S. Leeder, T.E. Klein, K.E. Caudle, et al., Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450 2D6 genotype and codeine therapy: 2014 update, Clin. Pharmacol. Ther. 95 (2014) 376–382.

[15] M.V. Relling, M. Schwab, M. Whirl-Carrillo, G. Suarez-Kurtz, C. H. Pui, C.M. Stein, et al., Clinical Pharmacogenetics Implementation Consortium guideline for thiopurine dosing based on TPMT and NUDT genotypes: 2018 update, Clin. Pharmacol. Ther. (2019) 1095–1105.

[16] A.H. Wu, Genotype and phenotype concordance for pharmacogenetic tests through proficiency survey testing, Arch. Pathol. Lab. Med. 137 (2013) 1232–1236.

Self-assessment questions (True/False)

1. Pharmacogenetic testing can always determine the specific drug that will benefit a patient most.
2. Pharmacogenetic testing prevents the need for therapeutic drug monitoring.
3. *HLA-B* alleles can predict an increased risk of hypersensitivity reactions to certain drugs.
4. Guidelines for implementation of pharmacogenetic testing describe the specific gene variants that should be targeted by a laboratory test.
5. Copy number variations in *CYP2D6* could influence the drug response phenotype.
6. A drug that is activated by metabolism may be contraindicated in a person that has inherited no function genetic variants for the primary enzyme involved in metabolism of that drug.
7. A person that inherits a loss of function variant in *TPMT* should not receive thiopurine drugs.
8. Star (*) alleles are commonly used to describe alleles for pharmacogenes.
9. Pharmacogenetic variants are rare in most populations.
10. Application of pharmacogenetic variants to clinical care requires consideration of drug interactions.

Answers

1. False
2. False
3. True
4. False
5. True
6. True
7. False
8. True
9. False
10. True

Chapter 54

Infectious diseases

Esther Babady

Clinical Microbiology Service, Department of Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, United States

Learning objectives

After reviewing this chapter, the reader will be able to:

- Describe clinical syndromes caused by herpesviruses, and advantages and limitations of each laboratory testing method.
- Describe clinical syndromes caused by hepatitis viruses, and advantages and limitations of each laboratory testing method.
- Describe clinical syndromes caused by HIV, and advantages and limitations of each laboratory testing method.

Herpesviruses

Herpesviruses are double-stranded, linear DNA viruses classified in the *Herpesviridae* family. Eight herpesviruses, subdivided into three subfamilies, cause infections in humans (Table 54.1). The *Alphaherpesvirinae* subfamily includes three viruses: herpes simplex virus 1 (HSV-1), herpes simplex virus 2 (HSV-2), and varicella-zoster virus (VZV). Three viruses make up the *Betaherpesvirinae* subfamily: cytomegalovirus (CMV), human herpesvirus 6, and human herpesvirus 7. The third subfamily, the *Gammaherpesvirinae*, includes Epstein–Barr virus (EBV) and human herpesvirus 8. Herpesviruses are responsible for a wide range of clinical syndromes ranging from asymptomatic infections to severe disease with devastating sequelae, particularly in the very young and in immunocompromised hosts.

Herpes simplex viruses

Herpes simplex viruses (HSVs), also known as human herpesviruses, include two antigenically different strains: HSV-1 and HSV-2. Entry of HSV-1 and HSV-2 into the host cells is facilitated through the interactions of viral surface glycoproteins to specific cell receptors. HSV viral particles are transported from the entry site to the sensory neurons by retrograde transport along the peripheral nerves, where they invade sensory neurons. Following primary infections, HSV-1 and HSV-2 establish latency in the dorsal root ganglia, with reactivation occurring due to immunosuppression or other stressors [1,2].

HSV-1 is transmitted primarily through contact with oral secretions and lesions. HSV-1 may infect any mucocutaneous surface, resulting in either an asymptomatic, subclinical infection or primarily in orolabial herpes, commonly known as "cold sores." HSV-1 keratitis is the

TABLE 54.1 Herpesviruses classification.

Viruses	Subfamily	Site of latency	Example of syndromes
Herpes simplex virus 1	*Alphaherpesvirinae*	Sensory neurons	Gingivostomatitis
Herpes simplex virus 2	*Alphaherpesvirinae*	Sensory neurons	Neonatal herpes
Varicella-zoster virus	*Alphaherpesvirinae*	Sensory neurons	Varicella (chickenpox)
Epstein–Barr virus	*Gammaherpesvirinae*	Lymphocytes	Infectious mononucleosis
Cytomegalovirus	*Betaherpesvirinae*	Tissues Endothelia cells Leukocytes	Pneumonitis
Human herpes virus 6	*Betaherpesvirinae*	Lymphocytes	Roseola
Human herpes virus 7	*Betaherpesvirinae*	Lymphocytes	Roseola
Human herpes virus 8	*Gammaherpesvirinae*	Lymphocytes	Kaposi sarcoma

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00054-5

most common cause of infectious corneal blindness in the United States. Other clinical syndromes associated with HSV-1 infections include esophagitis, pneumonitis, hepatitis, encephalitis, and myelitis. Although infections caused by HSV-1 were traditionally thought to occur only above the waistline, current data show that HSV-1 is responsible for ~25% of genital herpes ulcers [1].

HSV-2 infections result from direct contact with genital secretions of individuals shedding the virus, including sexual contact or through perinatal transmission during vaginal delivery. HSV-2 is the most common cause of genital ulcers worldwide with >500 million people infected. A recent report from the World Health Organization (WHO) estimates that neonatal herpes occurs in 10 per 100,000 live births, equivalent to an overall burden of 14,000 cases per year [3]. Neonatal herpes can result in serious long-term neurologic sequelae if not diagnosed and treated quickly. Risk factors for transmission of HSV-2 from pregnant women to neonates include primary infection during pregnancy, seropositivity status of the mother, and delivery methods. HSV-2 causes clinical syndromes similar to those described for HSV-1, as well as aseptic meningitis, which is a complication of genital herpes that presents in a small percentage of patients [2]. Current antiviral options for HSV infections include intravenous acyclovir for neonatal herpes and meningitis and oral acyclovir, famciclovir, or valacyclovir for genital herpes [4–6].

Laboratory testing

Although the viruses share structural similarities, HSV-1 and HSV-2 differ by up to 50% in their genome sequences, allowing for easy discrimination by molecular methods. Differentiation of the two serotypes using commercial immunoassays is more challenging, as both viruses have several antigens that share similar epitopes. Commercial, type-specific antibody assays that differentiate between HSV-1 and HSV-2 anti-IgG are available. Serological testing may be useful in providing a diagnosis in subclinical or asymptomatic infections, where the virus or viral DNA may be undetectable, and for rapid evaluation of patients presenting with clinical manifestations of sexually transmitted diseases [2]. The importance of distinguishing between HSV-1 and HSV-2 infections using type-specific serological assays includes the following: (1) determination of the serostatus of pregnant women to evaluate risk of acquiring HSV infections from a serodiscordant partner; and (2) the identification of asymptomatic herpes (i.e., in the absence of lesions) to distinguish between genital and oral herpes [7,8].

Serological HSV assays have several limitations. First, they have limited utility in diagnosing acute infections in neonates (i.e., inability to differentiate from mother IgG) or other patient populations, including immunocompromised patients [9,10]. Second, while genital herpes is primarily due to HSV-2 and oral herpes is associated with HSV-1, serological assays cannot identify the site of infection. Third, the levels of IgM may not be detectable early in the course of the infection. Finally, false-positives due to cross-reactivity between the two viruses may occur, most commonly in low prevalence populations [2].

Both direct fluorescent antibody (DFA) and indirect fluorescent antibody (IFA) assays are commercially available for detection of HSV antigens in clinical specimens. Type-specific monoclonal antibodies are used to differentiate between HSV-1 and HSV-2. In addition, DFA/IFA may be used to confirm a cytopathic effect, which is characteristic of HSV detected in viral cell culture. The test principles are based on the visual detection of virally infected cells, fixed on a glass slide, using fluorescein-labeled monoclonal antibodies against HSV-1/2 antigens [11]. These assays are manual and include several steps (centrifugation, incubation, and reading); however, they are associated with relatively rapid turnaround times (TATs) of ~2 hours. However, DFA/IFA testing performed directly on clinical samples is not recommended for routine diagnosis of acute HSV infections due in part to their lower sensitivity when compared with viral culture and molecular tests [2,11]. When performed, many laboratories will reflex negative DFA/IFA tests to either viral cultures or molecular tests.

Nucleic acid amplification test (NAT) or molecular tests are the most sensitive method for the diagnosis of active HSV infections. Several commercial assays are FDA-cleared or approved for the qualitative detection of HSV-1 and HSV-2 in cutaneous and mucocutaneous lesions (Table 54.2). In addition to detecting viral nucleic acids, these molecular tests can differentiate between HSV-1 and HSV-2. The sensitivity and specificity of molecular tests range between 95% and 100% when compared with viral culture, and assays have laboratory TATs ranging from 20 minutes to <4 hours. This is markedly improved when compared against the laboratory TAT of viral cultures, which are on the order of days. Detection of HSV DNA alone, however, should be interpreted with caution, since asymptomatic shedding has been reported and its significance in transmission of HSV infection is not well understood [2].

Aside from the detection of HSV in lesions, molecular testing is the method of choice for the diagnosis of aseptic meningitis/encephalitis (M/E) caused by herpesviruses, via the detection of HSV-1/2 in cerebrospinal fluid (CSF) [12]. However, to date, only one real-time PCR assay, the Simplexa HSV 1 and 2 direct test (Focus Diagnostics, Inc.), is commercially available for detection of HSV in CSF. This test is a qualitative, real-time PCR assay that amplifies, detects, and differentiates between HSV-1 and

TABLE 54.2 Select diagnostic tests for herpesviruses.

Viruses	Method	FDA approved	Characteristics	Target	Intended use	Specimens
HSV-1/2	IA	Yes	Qualitative	IgG	Immune status Differentiate between HSV-1 and HSV-2	Serum Plasma
	IA	Yes	Qualitative	IgM	Diagnosis	Serum
	NAT	Yes	Qualitative	Viral DNA	Diagnosis	Cutaneous mucocutaneous CSF
	NAT	No	Qualitative	Viral DNA	Diagnosis	Any specimen types
CMV	IA	Yes	Qualitative	IgG	Immune status	Serum Plasma
	IA	Yes	Qualitative	IgM	Diagnosis	Serum
	IA	Yes	Qualitative	IgM and IgG	Diagnosis	Serum
	IA	Yes	Qualitative	pp65	Diagnosis	Leukocytes
	NAT	Yes	Quantitative	Viral DNA	Diagnosis Monitoring	Plasma Whole blood
	NAT	Yes	Qualitative Quantitative	Viral DNA	Diagnosis	Any specimen types
EBV	IA	Yes	Qualitative	Heterophile antibodies VCA IgG VCA IgM VCA IgA EBNA-1 IgG NA-1 IgG EA-D Ig	Diagnosis	Serum
	NAT	No	Quantitative	Viral DNA	Diagnosis Monitoring	Plasma Whole blood
	NAT	No	Qualitative	Viral DNA	Diagnosis	Any specimen types

CMV, Cytomegalovirus; *CSF*, cerebrospinal fluid; *EA*, early antigen; *EBNA*, Epstein–Barr virus nuclear antigen.; *EBV*, Epstein–Barr virus; *HSV-1*, herpes simplex virus; *HSV-2*, herpes simplex virus 2; *IA*, immunoassay; *IVD*, in vitro diagnostics; *NAT*, nucleic acid amplification test; *VCA*, viral capsid antigen.

HSV-2 in CSF samples without nucleic acid extraction prior to amplification. The sensitivity and specificity of this approach, when compared against a laboratory-developed test (LDT) based on Roche analyte-specific reagents (ASRs), were 96.2% and 97.9%, respectively [13]. An FDA-cleared multiplexed molecular panel, The BioFire M/E panel, is commercially available for the rapid detection of common pathogens associated with infectious M/E, including HSV-1 and HSV-2. In one study, the percent agreement between the M/E panels for the detection of HSV-1 and HSV-2 compared against singleplex HSV-1 and HSV-2 were 82.6% and 96.0%, respectively [14].

Human cytomegalovirus

Human CMV, also known as human herpesvirus 5, is one of the most common infections worldwide, with an estimated age-adjusted seroprevalence of 50.4% in the United States. Infections caused by CMV range from subclinical or asymptomatic infections to severe diseases that may occur in neonates and immunosuppressed hosts, including solid-organ transplant (SOT) and hematopoietic stem cell transplant (HSCT) recipients.

Transmission of CMV occurs following exposure to body fluids from infected individuals. Upon primary infection, CMV viral particles enter host cells by binding to cell surface receptors via viral glycoproteins. Internalization of viral particles occurs either through fusion or endocytosis. Viral replication occurs in the host nucleus followed by virion assembly inside the Golgi apparatus prior to extracellular release and propagation. CMV has wide cell tropism, as such infections result in a variety of syndromes. Lifelong latency, where the virus does not replicate, is established primarily in myeloid cells (i.e., monocytes), and reactivation can occur at any time, particularly postimmune trauma (i.e., inflammation and immunosuppression).

The following modes of CMV transmission have been described: vertical transmission from mother to child (intrauterine or perinatally), horizontal transmission following close contact, such as in daycare centers or sexual contacts, and via blood transfusions or organ transplantations. Primary infection in healthy hosts is often asymptomatic, although, in some cases, patients may develop either a mononucleosis syndrome similar in presentation (fatigue and fever) to infections caused by the EBV. In rare instances, more serious complications, such as arthralgia, myalgia, and aseptic meningitis, have been observed [15].

CMV remains the most common congenital infection in the Western world. Current approaches to prevention of congenital CMV infections include passive immunization of seronegative mothers through administration of CMV hyperimmune globulin, thereby reducing potential exposure to infected individuals. Although most congenital infections are asymptomatic, approximately 10%–15% of neonates develop symptomatic congenital CMV as a result of primary infection in utero. Of these, 90% will develop sequelae, including cognitive deficits, as well as hearing and visual loss. The degree of severity of congenital CMV sequelae is highest if transmission occurs during the first trimester and less severe if acquired perinatally [15,16].

CMV disease, defined as the presence of detectable CMV DNA or antigen and clinical symptoms, is a significant cause of morbidity and mortality in both SOT and HSCT recipients [17]. Risk factors associated with the development of CMV disease in SOT include the CMV serostatus of both the donor (D) and the recipient (R), with the highest risk seen for the donor positive/recipient negative (D+/R−) pair, the type of organ transplanted and the type and degree of immunosuppression posttransplantation. Current management strategies for transplant patients include antiviral prophylaxis and preemptive therapy, both of which have resulted in decreased incidence of early (i.e., first 4 months posttransplant) CMV disease. However, late CMV infections still occur, with gastrointestinal disease being the most common presentation in most transplant groups [18].

Protection against CMV infection is mediated through T-cell immunity, particularly involving both CD8 and CD4 T-lymphocytes. Highly immunosuppressed, CMV seropositive HSCT recipients, including those receiving T-cell depleted or unrelated donor grafts, are particularly at high risk of developing CMV disease. In addition, both autologous and allogeneic HSCT treated with high-dose corticosteroids are at increased risk for CMV infection. The most serious CMV disease manifestation in HSCT recipients is CMV pneumonia, which has a significantly high mortality rate, even with current antiviral management approaches [19,20].

Laboratory testing

Several serologic assays are available for the diagnosis of CMV infections (Table 54.2). These assays provide indirect evidence of recent or past infections and are used primarily to establish serostatus. Diagnosis of acute CMV infection by serology requires the collection of paired serum samples with the first sample collected soon after symptoms appear, and a second sample collected 2–4 weeks following the primary infection. A fourfold increase in CMV IgG titers between the samples collected at these time points strongly suggests a recent or acute infection [21]. Alternatively, performance of single CMV IgM or CMV IgG plus IgG avidity testing can be used to diagnose an acute infection. In general, the utility of serological testing for acute infections is relatively low, especially in cases where a follow-up, convalescent serum sample (following resolution of symptoms) cannot be collected, or in posttransplant patients who may have delayed immune recovery. CMV serology has a high prognostic value and is useful in determining CMV status of expectant mothers and during pretransplant evaluations of donors and recipients to determine risk of transmitting virus to a child or acquiring posttransplant infections, respectively.

More direct methods to the diagnosis of acute CMV infection are based on the detection of CMV antigens or nucleic acids. Detection of CMV antigens in blood (antigenemia) is used for both diagnosis and monitoring of patients with CMV infections. The most common antigen measured is pp65, a 65-kDa phosphoprotein found in infected leukocytes and endothelial cells [22]. Detection of antigenemia requires collection of 5–10 mL of whole blood followed by isolation of leukocytes and fixation on glass slides. Monoclonal CMV antibodies are then added to form a complex with the pp65 antigens present in fixed leukocytes. A second antibody, conjugated to fluorescein isothiocyanate, binds to the antibody–antigen complex. Positive cells fluoresce green and can be observed and counted under a fluorescent microscope. Although the assay time is relatively short (~2 hours), and the assay does not require specialized laboratory equipment, except for a fluorescent microscope, antigenemia has several limitations: (1) leukocytes have to be isolated soon after collection for maximum sensitivity; (2) the assay is labor-intensive with several steps (including centrifugation and multiple wash steps) and long hands-on time; and (3) a minimum number of cells (e.g., 150,000–200,000 cells) have to be obtained for the test to be valid, which limits the utility of the assay in neutropenic patients. Furthermore, interpretation of the test is subjective, and there is no standardization across assays and laboratories. Various thresholds (e.g., number of pp65-positive cells/50,000 leukocytes) have been established by individual

centers to distinguish asymptomatic shedding from active infection, to initiate or discontinue antiviral therapy, and to predict the risk for acquiring CMV diseases [22]. However, thresholds are not standardized across transplant centers, and this method has been phased out in most clinical laboratories.

In most centers, CMV infections and CMV diseases are often diagnosed using molecular assays due to their rapid TATs, increased sensitivities, and objective interpretation of results. Because CMV infections do not always predict development of CMV disease, the increased sensitivity of molecular assays may also present a challenge when determining the significance of a positive result. The most common amplification method used for CMV detection is real-time PCR, including qualitative and quantitative molecular methods. Qualitative real-time PCR assays are primarily used for the detection of CMV DNA in various body fluids (e.g., CSF and bronchial washings) and tissues. All current assays are LDTs based on either commercially available ASRs or user-developed primers and probes [23]. The lower limit of detection (LOD) among molecular assays can vary as much as 10-fold, and such information should be considered carefully during assay selection and implementation. High sensitivity may be required for diagnosis of central nervous system infections, in which any positive result would be considered diagnostic of CMV M/E and requires immediate treatment, while a higher limit of detection may be acceptable for a respiratory specimen, where asymptomatic shedding is more common and additional clinical information would be required to diagnose CMV pneumonia.

Quantitative real-time PCR assays are an integral part of the clinical management of transplant patients. Early detection of CMV by quantitative real-time PCR often results in more frequent monitoring of viral loads. Once a preestablished viral threshold is reached (depending on patient population), appropriate antiviral therapy may be implemented to prevent development and progression of CMV disease. This approach, referred to as preemptive therapy, has been shown to result in less drug toxicity compared with a prophylaxis approach, where antiviral drugs are administered prior to detection of CMV reactivation.

Until recently, most quantitative real-time PCR assays used were LDTs, which showed a high degree of variability across centers and made standardization challenging. This variability is explained in part by differences in genes targeted for amplification, sample types used, nucleic acid extraction and amplification methodologies, and units used to express results. Recently, two developments occurred that resulted in improved standardization of CMV results: (1) the release of the first WHO CMV international standards; and (2) the approval by the FDA of several CMV quantitative molecular assays. All FDA-cleared assays have traceability to the first WHO International Standard for Human Cytomegalovirus. In one multicenter, international study, the Roche quantitative assay was shown to perform consistently well across laboratories when testing was performed on a standardized CMV panel. However, variability between each individual LDT and the Roche assay was observed when using a set of plasma samples from immunosuppressed patients [24]. Several studies have reported data on head-to-head comparisons of FDA-cleared molecular tests for CMV viral loads. In general, the assays correlate well (r value > 0.8), although differences were observed at the lower end of the linear range for assays with LOD (e.g., Abbott real-time CMV vs the Roche CAP/CTM assay) [25,26].

Epstein–Barr virus

EBV, also known as human herpesvirus 4, is transmitted through contact with contaminated oral secretions, blood products, or transplanted cells and/or organs. Similar to CMV, initial exposure to EBV occurs between early childhood and adolescence [27]. Lytic replication of EBV occurs in both B-lymphocytes and epithelial cells, followed by establishment of a latent viral state in B-lymphocytes. Two of the most recognized syndromes caused by EBV are infectious mononucleosis (IM) and posttransplant lymphoproliferative disorders (PTLDs). In immunocompetent hosts, primary infection with EBV either may be asymptomatic or result in IM. The illness is mild and characterized by nonspecific symptoms, such as fever, headache, sore throat, and fatigue; the most commonly characteristic clinical triad includes pharyngitis, fever, and cervical lymphadenopathy [27]. PTLDs result from either reactivation or primary infection with EBV in immunosuppressed hosts (e.g., SOT and HSCT recipients) following loss of T-cell functions [28]. Manifestations of PTLDs vary from benign to severe diseases related to the oncogenic properties of EBV, with EBV-associated lymphomas (e.g., Burkitt's lymphoma) and carcinomas (e.g., nasopharyngeal carcinoma) representing the severe end of the spectrum [28,29].

Laboratory testing

Heterophile antibodies.

Diagnosis of EBV infections by serology can be complex and is based on the detection of several immunoglobulins including both EBV nonspecific antibodies and EBV-specific antibodies [30]. Heterophile antibodies are produced in response to antigens produced during EBV IM (EBV heterophile antigens or Paul–Bunnell antigens) or as a result of serum sickness (type III hypersensitivity reaction caused by proteins present in certain medications)

or rheumatoid factors (non-EBV heterophile antigens or Forssman antigens). These molecules are nonspecific for EBV-derived antigens. Traditionally, detection of heterophile antibodies relied on their ability to bind and agglutinate erythrocytes of various animal species (sheep, goat, and horse). To increase assay specificity for EBV heterophile antibodies, serum samples were first incubated with guinea pig kidney cells to adsorb and remove any Forssman heterophilic antibodies present in the sample. Today, most agglutination tests (e.g., monospot test) use polystyrene latex beads coated with Paul–Bunnell antigens instead of animal erythrocytes. Precipitation of latex beads with a patient serum sample within 10 minutes suggests the presence of EBV heterophile antibodies. In rare instances, false-positive results may occur in patients with lymphoma or autoimmune diseases, and false-negative results may occur in pediatric patients.

Anti-EBV antibodies.

Detection of an antibody response to the following EBV viral antigens is used to diagnose various stages of EBV infections: the viral capsid antigen (VCA), early antigen (EA), and Epstein–Barr virus nuclear antigens (EBNAs) (Table 54.3). Antibodies (IgM and IgG) to the VCA appear first, often within 1–4 weeks of the primary infection. While VCA IgG remains detectable for life, VCA IgM levels decrease and disappear within the first 8 weeks; the loss of VCA IgM is followed by the appearance of VCA IgA. In about 5% of cases, VCA IgM may persist, resulting in aberrant serological profiles that require testing of additional markers for proper result interpretation (Table 54.3). Anti-EA IgG also appears early during the infection life cycle and may persist for several months. EBNA-1 IgG appears last and persists for several months to years. Acute EBV infections are characterized by the detection of VCA IgM with or without IgG, while past infection shows VCA IgG and EBNA-1 IgG only. However, EBV serology profile interpretation may be challenging and requires supplemental testing, including IgG avidity testing, which is used to evaluate the maturity of IgG, with recent infections characterized by low avidity (weak binding and immature IgG) and past infections characterized by high affinity (strong binding and mature IgG). Other useful supplemental testing includes heterophile antibodies and NAT [30].

NAT approaches are useful for the detection of EBV DNA in a wide range of specimen types and for monitoring of EBV reactivation in blood samples (whole blood, serum, plasma, or peripheral blood monocyte) of high-risk patients (including SOT and HSCT recipients) to prevent PTLD [31] (Table 54.2). Because EBV may reactivate without causing symptoms, quantitative NAT assays are most useful in distinguishing significant viral loads from background levels, although no standard threshold has yet been established. High viral loads are often detected in patients with PTLD, and changes in viral loads of >0.5 log copies/mL between consecutive samples are traditionally considered significant. However, to date, there are no FDA-cleared molecular EBV NAT assays. All molecular EBV assays are LDTs, which may result in increased variability across laboratories and do not allow for results to be easily compared across centers. Sources of variability in the use of LDTs include viral target (e.g., single copy vs multicopies target), sample types (whole blood vs plasma), nucleic acid extraction methods (manual vs automated), amplification platforms, calibrators and standards used, and reporting units (e.g., copies/mL vs copies/mg) [32]. To standardize the wide range of NAT methods available, the first WHO International Standards for EBV were developed in 2011 and became available for calibration of secondary reference materials to international unit (IU). In a recent study, the use of the EBV WHO standards to convert viral loads from copies/mL to IU/mL resulted in lower EBV values and reduced variability in viral load measurements [33]. These data are encouraging as laboratories continue to aim for standardization of EBV viral loads and establishment of universally clinically significant EBV viral load thresholds.

Hepatitis viruses

Hepatitis viruses are a diverse group of viruses that include both DNA and RNA viruses and are responsible for causing viral inflammation of the liver. These viruses

TABLE 54.3 Interpretation of Epstein–Barr virus serological profile.

Marker	Susceptible host	Acute infection	Past infection	Chronic infection	Inconclusive[a]		
VCA IgM	−	+	−	+/−	−	+	−
VCA IgG	−	+/−	+	+	+	+	−
EBNA IgG	−	−	+	+	−	+	+

EBNA, Epstein–Barr virus nuclear antigen; *VCA*, viral capsid antigen.
[a]*Supplemental testing required: EBV DNA test, IgG avidity, heterophile antibodies, and early-antigen antibodies.*

TABLE 54.4 Hepatitis viruses.

Viruses	Family	Genus	Genome	Transmission	Chronic infection
Hepatitis A	*Picornaviridae*	*Hepatovirus*	Linear, ssRNA	Fecal–oral	No
Hepatitis B	*Hepadnaviridae*	*Orthohepadnavirus*	Circular, partially dsDNA	Sexual Blood-borne Transplant Vertical	Yes
Hepatitis C	*Flaviviridae*	*Hepacivirus*	Linear, ssRNA	Sexual Blood-borne Transplant	Yes
Hepatitis D	Unassigned	*Deltavirus*	Circular, ssRNA	Blood-borne	Yes
Hepatitis E	*Herpesviridae*	*Herpesvirus*	Linear, ssRNA	Fecal–oral	Yes

differ in their genome composition, structure, mode of transmission and associated syndromes, and their propensity for causing chronic infections. All hepatitis viruses can cause jaundice and elevations in serum alanine aminotransferase (ALT) or aspartate aminotransferase (AST) activity measurements. In general, acute hepatitis is clinically defined as an acute illness with discrete onset of symptoms (e.g., fever, headache, malaise, anorexia, nausea, vomiting, diarrhea, and abdominal pain) and jaundice, typically in the presence of elevated liver enzyme (AST and ALT) measurements. Laboratory testing is required to confirm viral hepatitis and correctly identify the responsible virus.

In addition to the five major hepatitis viruses (Table 54.4) discussed in this section, two other hepatitis viruses, hepatitis G virus and GB virus type C, have been described. Transmission of these two viruses may occur through blood transfusion or intravenous drug use. However, their prevalence and clinical significance are not well understood [34].

Hepatitis A

Hepatitis A virus (HAV) is a single-stranded RNA virus belonging to the *Picornaviridae* family along with other gastrointestinal viruses, such as enteroviruses. As a nonenveloped virus, HAV can survive in the environment for several months, requiring high temperatures ($>185°F$) and chlorine for inactivation. HAV is transmitted through ingestion of contaminated food or water and causes gastroenteritis symptoms, such as nausea and diarrhea. The incubation period of HAV is ~ 28 days, and symptoms resolve rapidly, usually within several weeks. Unlike other hepatitis viruses, there have not been any reports of HAV causing chronic infections [35]. Vaccination for HAV was introduced in the United States in 1996 and is recommended by the Advisory Committee on Immunization Practice as part of the routine childhood vaccination schedule, with most children receiving vaccination before the age of 2 years. In a 2007 report, the rate of HAV infections in the United States was reported at 1.0 case per 100,000 persons, a 92% decrease compared with the 1995 rate of 12.0 case per 100,000 population [36].

Laboratory testing

Symptoms caused by HAV are indistinguishable from infections caused by other hepatitis viruses. In addition to clinical symptoms described above, acute HAV infection requires laboratory testing derived primarily from the serologic detection of antibodies (IgM and IgG) to HAV capsid proteins. HAV IgM antibodies appear 5–10 days prior to the onset of symptoms, and become undetectable within 6 months of exposure. IgG antibodies appear 4–8 weeks after infection and remain present for the lifetime of the patient. Commercially available serologic assays are designed to detect total anti-HAV (IgG and IgM), and, if positive, a reflex test for anti-HAV IgM alone can be performed to diagnose acute infections. Although enzyme immunoassays (EIAs) for anti-HAV IgM are sensitive, false-positive results have been reported in various situations, including in patients with autoimmune diseases, in the presence of high titers of rheumatoid factor, and in patients with resolved HAV infections. These situations underscore the importance of testing for HAV in the appropriate clinical context [37].

Hepatitis B

Hepatitis B virus (HBV) is a small, circular, partially double-stranded DNA virus from the *Hepadnaviridae* family. The HBV virions exist as both infectious and noninfectious particles. The noninfectious particles are spherical or tubular-enveloped virions with embedded hepatitis B surface antigen (HBsAg), but no other viral

components. The infectious viral particles, known as Dane particles, are complete virions with HBsAg in the envelope, a viral genome, and both the HBV core antigen and the hepatitis e antigen contained within the virion [38]. Up to 10 HBV genotypes (genotypes A−J) and >30 subtypes have been described to date. HBV genotypes differ by >8% in their genome sequences, while subtypes differ by 4%−8%. HBV has a worldwide distribution with genotypes showing variability across geographic regions. In the United States, genotypes A, B, and C are the most prevalent with reported rates of 35%, 22%, and 31%, respectively [39].

HBV is the leading cause of viral hepatitis, with over 2 billion infected people worldwide and >350 million with chronic infections. In 2017, an estimated 22,100 cases of acute HBV occurred in the United States, a rate that has remained stable over the last several years due in part to successful immunization efforts and better surveillance. This is despite a slight increase in incidence, which has been associated with the recent opioid crisis [40,41]. Transmission of HBV occurs through direct contact with infected body fluids. In high prevalence regions (i.e., >8%), transmission of HBV most often occurs perinatally, from infected mothers to neonates during delivery. In low endemicity regions, transmission occurs primarily through percutaneous and sexual routes, or from close contact.

Symptoms of HBV infections are similar to those caused by other viral hepatitis agents, with malaise, fever, nausea, and possible jaundice being reported. While acute HBV infection is self-limited in most individuals, progression to chronic HBV does occur in a subset of infected patients. Risk factors for developing chronic HBV infections include host immune function and age when exposure occurs; up to 50% of children develop chronic HBV infection when exposed between the ages of 1 and 5. This number drops to ~10% if older children and adults become virally infected [42]. Chronic HBV infection proceeds in three distinct phases: an immune-tolerant phase, an immune-active phase, and an inactive carrier phase. The immune-tolerant phase occurs in infants who acquired the virus perinatally and is characterized by active HBV replication, but no disease or symptoms. The immune-active phase occurs when the HBV is acquired from person to person and is characterized by active disease with positive HBV serologies and liver inflammation. Patients may progress from the immune-tolerant phase to the immune-active phase over several decades. The inactive carrier phase is a resolution phase, where patients have recovered from acute infection and have no or little active disease (i.e., no liver abnormalities). Patients in the inactive carrier phase may reactivate the virus in cases of stress or immunosuppression. Complications of chronic HBV infections include liver cirrhosis and hepatocellular carcinomas (HCCs), which carry high morbidity and mortality rates [43].

Laboratory testing

Diagnosis of acute and chronic HBV infections is based on the detection of various serological markers as summarized in Table 54.5 [44]. The incubation period following exposure to HBV ranges from 6 weeks to 6 months. Several commercial assays are FDA-cleared and available for the detection of HBV antigens and antibodies. Most methods used are based on EIA and CLIA (Table 54.6). Laboratory testing via immunoassays is used for screening, diagnosis, and confirmation of HBV immunity.

Hepatitis B surface antigen and antibodies

Acute HBV infection is characterized by high titers of HBsAg, which appear within 3−4 weeks postexposure, peaks at ~12 weeks, and becomes undetectable after 4−6 months. Detection of HBsAg often requires confirmation, especially when the signal-to-cutoff threshold is low, to prevent the reporting of false-positive results.

TABLE 54.5 Interpretation of hepatitis B serological pattern.

Marker	Incubation period	Acute HBV infection	Past HBV infection	Chronic HBV infection	Occult HBV infection	HBV vaccination
HBsAg	+	+	−	+	−	−
HBeAg	+	+	−	+/−	+/−	−
Anti-HBc IgM	−	+	−	−	+/−	−
Anti-HBc total	−	+	+	+	+	−
Anti-HBe	−	−	+/−	+/−	−	−
Anti-HBs	−	−	−	−	−	+

anti-HBc, Hepatitis B core antibody; HBsAg, hepatitis B surface antigen; HBV, hepatitis B virus.

TABLE 54.6 FDA-approved hepatitis viruses and HIV-1 FDA-approved immunoassays.

Manufacturer	Test name	Method[a]	Target	Generation
Abbott Diagnostics	ARCHITECT HBsAg Qualitative, HBsAg Qualitative Confirmatory, Core, Core IgM, AUSab	CLMIA	HBV	N/A
	AxSYM Core 2.0., Core M 2.0., HBsAg, HBsAg confirmatory	MEIA	HBV	N/A
	Architect Anti-HCV	CLMIA	HCV	Third
	AxSYM Anti-HCV	MEIA	HCV	Third
	Architect HIV Ag/Ab Combo	CLMIA	HIV-1 HIV-2 p24	Fourth
	Alere Determine HIV-1/2 Ag/Ab Combo	LFIA	HIV-1 HIV-2 p24	Fourth; waived
bioLytical Laboratories, Inc.	INSTI HIV-1/HIV-2 Rapid Antibody Test	LFIA	HIV-1 HIV-2	Third; waived
Bio-Rad Laboratories, Inc.	GS HBsAg, MONOLISA Anti-HBc IgM, Anti-HBc, Anti-HBs	EIA	HBV	N/A
	Geenius HIV 1/2 Supplemental Assay	ICA	HIV-1 HIV-2	Third; waived
	Multispot HIV-1/HIV-2 Rapid Test	ICA	HIV-1 HIV-2	Third
	GS HIV Combo Ag/Ab	EIA	HIV-1 HIV-2 p24	Fourth
	BioPlex 2200 System HIV Ag−Ab	EIA	HIV-1 HIV-2 p24	Fifth
Chembio Diagnostics	DPP HIV 1/2 Assay	LFIA	HIV-1 HIV-2	Third
	SURE CHECK HIV 1/2 Assay	LFIA	HIV-1 HIV-2	Third
	HIV 1/2 STAT-PAK Assay	LFIA	HIV-1 HIV-2	Third
Ortho Clinical Diagnostics	VITROS Anti-HBe, HBeAg, HBsAg, Anti-HBs, Anti-HBc,	CLIA	HBV	N/A
	Vitros Anti-HCV	CLIA	HCV	Third
	Vitros Anti-HIV 1 + 2	CLIA	HIV-1 HIV-2	Third
Roche Diagnostics	Elecsys Anti-HBc, Anti-HBc IgM, HBsAg, Anti-HBs, HBeAg	ECLIAECLIA	HBV	N/AN/A
	Elecsys Anti-HCV		HCV	
Siemens Healthcare Diagnostics	Advia Centaur Anti-HBe, Anti-HBs Anti-HBc IgM, Ant-HBc Total HBeAg, HBsAg, HBsAg Confirmatory	CLIA	HBV	N/A
	IMMULITE Anti-HBs, Anti-HBc IgM, Anti-HBc Total, Anti-HBs Quantitative, HBsAg, HBsAg Confirmatory	CLIA	HBV	N/A
	Advia Centaur HCV	CLIA	HCV	Third
	Advia Centaur HIV 1/O/2	CLIA	HIV-1 HIV-1, Group O HIV-2	Third
OraSure Technologies, Inc.	OraQuick HCV	LFIA	HCV	Third
	OraQuick *ADVANCE* Rapid HIV-1/2 Antibody	LFIA	HIV-1 HIV-2	Third
Trinity Biotech	Uni-Gold Recombigen HIV-1/2	LFIA	HIV-1 HIV-2	Third

anti-HBc, Hepatitis B core antibody; *Anti-HBe*, hepatitis B e antibody; *HBsAg*, hepatitis B surface antigen; *HBV*, hepatitis B virus; *HCV*, hepatitis C virus.
[a]CLMIA, *Chemiluminescent microparticle immunoassay;* MEIA, *microparticle enzyme immunoassay;* CLIA, *chemiluminescent immunoassay;* EIA, *enzyme immunoassay;* ECLIA, *electrochemiluminescence assay;* LFIA, *lateral flow immunoassay;* ICA, *immunochromatograph assay*.

Algorithms for samples requiring reflex to confirmatory testing is usually established by manufacturers and performed per package insert instructions. Confirmation is done via a neutralization assay, in which the positive patient serum sample is pretreated with or without high titers of anti-HBsAg. The resulting antigen−antibody

complex effectively neutralizes the HBsAg, if present. A decrease of at least 50% in the intensity of the signal between the neutralized and nonneutralized samples is considered confirmation of HBsAg presence. False-negative HBsAg results may occur due to genetic variants that result in structural changes in the "a" determinant, the major antigenic determinant of HBV, which is recognized by most of the commercial monoclonal antibodies. These HBsAg variants are designated "escape mutants" and may cause occult HBV infection, in which HBV DNA is detected in the liver with or without detectable HBV DNA in the serum in an HBsAg negative patient.

Detection of hepatitis B surface antibodies (anti-HBs) reflects recovery from acute HBV infection and provides immunity against reinfection. In cases of natural infection and recovery, both the anti-HBs and total hepatitis B core antibody (anti-HBc) are detectable. Detection of anti-HBs alone indicates immunity acquired through hepatitis B vaccination or through passive immunization with hepatitis B immune globulin, a prophylaxis treatment to prevent HBV infection in certain patient populations (e.g., infants exposed perinatally or immunosuppressed patients).

Hepatitis B e antigen and antibodies

Active replication of HBV virus is measured through detection of either the hepatitis B e antigen (HBeAg) or HBV DNA. HBeAg, a soluble protein contained in the core of HBV virions, becomes detectable within 4 weeks of exposure and disappears around 12 weeks postinfection. The presence of HBeAg in serum indicates a high degree of infectivity. Similar to HBsAg detection, serum samples positive for HBeAg must be confirmed through a neutralization assay. Mutations in the core and precore regions of the HBV genome can result in virus that does not produce HBeAg. HBeAg-negative patients with chronic HBV have a higher risk of clinical complications. HBeAg can also be used to monitor treatment response. Detection of the anti-HBeAg reflects a decrease in infectivity and recovery from acute infection.

Hepatitis B core antigen and antibodies

The core antigen of HBV is an intracellular antigen, present in hepatocytes but not detectable in serum. Shortly, after the appearance of HBsAg, anti-HBc IgM appears and remains detectable for up to 8 months; this is followed by the appearance of anti-HBsAg. In chronic HBV infections, both the HBsAg and the total anti-HBc molecule remain detectable over the course of months to years. Antibodies to the core antigens are measured as either IgM only or total anti-HBc. A negative total anti-HBc suggests no exposure to HBV, while a positive total anti-HBc may reflect acute, chronic, or resolved infections, depending on the presence of other markers of HBV infection.

Hepatitis B DNA

Molecular assays for HBV DNA are used for diagnosis, as well as to establish pretreatment viral load baseline, monitor treatment response, and viral genotyping [45]. FDA-approved molecular assays are available for the detection of HBV DNA virus (Table 54.7). Detection of HBV DNA is especially useful to diagnose acute HBV infection in cases of occult HBV infections and in cases of negative HBeAg due to precore or core mutants. Quantitative HBV DNA NAT assays have sensitivities ranging from 10 to 29 IU/mL, with upper limits of quantitation of $>10^7$ IU/mL. Patients with viral loads of at least 2000 IU/mL (if HBeAg is negative) or 20,000 IU/ml (if HBeAg is positive) and elevated liver enzymes qualify for antiviral treatment.

Treatment of chronic HBV infections requires long-term therapy, ranging from 6 months to 36 months. Current treatment options include the use of nucleoside (e.g., lamivudine and entecavir) and nucleotide (e.g., tenofovir) analogs with immune-modulating drugs (pegylated interferon) [46]. Given the extended duration of treatment, HBV strains can develop mutations associated with resistance and treatment failure. Although not routinely performed, genotyping assays are useful in predicting response to therapy (genotype B often responds better) and in predicting risk of progression to liver diseases. Direct sequencing of HBV may be used to detect the presence of mutations known to confer resistance to antiviral agents used for treatment of chronic HBV. There are currently no FDA-cleared methods for genotyping, although two assays targeting the polymerase gene of HBV genome are commercially available for use.

Hepatitis C

Hepatitis C virus (HCV) is a single-stranded, linear RNA virus belonging to the *Flaviviridae* family in the *Hepacivirus* genus. Other members of this family include arthropod viruses, such as West Nile virus and Dengue virus. HCV was discovered in 1989 and originally termed non-A, non-B hepatitis, as reagents against the known hepatitis viruses at the time were not reactive with this new entity. The HCV genome is ~9.6 kb and encodes 10 proteins, including six nonstructural (NS; e.g., NS2 and NS3) and four structural (a core protein, transmembrane protein, and two glycoproteins) proteins. Six HCV genotypes and >50 subtypes have been described to date. The genotypes have <60% sequence identity in the overall genome, while subtypes have >75%–85% identity in their core/E1 and NS5B sequences. The most common

TABLE 54.7 Hepatitis viruses and HIV-1 FDA-approved nucleic acid tests.

Manufacturer	Test name	Method	Characteristic	Linear range or LOD
Abbott	Realtime HBV	PCR	Quantitative	10−1,000,000,000 IU/mL
Molecular	Realtime HCV	RT-PCR	Quantitative	12−100,000,000 IU/mL
	Realtime HIV-1	RT-PCR	Quantitative	40−10,000,000 copies/mL
	Realtime CMV	PCR	Quantitative	50−1.56 × 100,000,000 IU/mL
Roche	Cobas HBV	PCR	Quantitative	10−1,000,000,000 IU/mL
Molecular	Cobas HCV	RT-PCR	Quantitative	15−100,000,000 IU/mL
	Cobas HIV	RT-PCR	Quantitative	20−10,000,000 copies/mL 60−10,000,000 copies/mL
	Cobas CMV	PCR	Quantitative	15−100,000,000 IU/mL
	CAP/CTM quantitative HBV v2.0	PCR	Quantitative	20−1.7 × 10,000,000 IU/mL
	CAP/CTM quantitative HCV v2.0	RT-PCR	Quantitative	15−100,000,000 IU/mL
	CAP/CTM quantitative HIV-1 v2.0	RT-PCR	Quantitative	20−10,000,000 copies/mL
	COBAS TaqMan HBV Test For Use With The High Pure System	PCR	Quantitative	29−1.1 × 10,000,000 IU/mL
	COBAS TaqMan HCV Test v2.0 For Use With The High Pure System	RT-PCR	Quantitative	25−300,000,000 IU/mL
	COBAS TaqMan HIV Test v2.0 For Use With The High Pure System	RT-PCR	Quantitative	34−10,000,000 copies/mL
	Cobas Amplicor HCV Monitor v2.0	RT-PCR	Quantitative	600−500,000 IU/mL
	Cobas Amplicor HCV test v2.0	RT-PCR	Qualitative	50−60 IU/mL
	Cobas Amplicor HIV Monitor v1.5	RT-PCR	Quantitative	50−75,000 copies/mL 400−750,000 copies/mL
	CAP/Cobas Amplicor HCV test v2.0	RT-PCR	Qualitative	50 IU/mL
	CAP/CTM HCV qualitative test v2.0	RT-PCR	Qualitative	15 IU/mL
Siemens	VERSANT HCV 3.0 Assay	bDNA	Quantitative	615−40,000,000 IU/mL
	VERSANT HCV RNA Qualitative Assay	TMA	Qualitative	5.3 IU/mL
	VERSANT HIV 3.0 Assay	bDNA	Quantitative	75−500,000 copies/mL
Hologic	Aptima HBV Quant Dx	TMA HPA	Quantitative	10−100,000,000 IU/mL
	Aptima HCV Quant Dx	TMA HPA	Quantitative	10−100,000,000 IU/mL
	Aptima HIV Quant Dx	TMA HPA	Quantitative	30−10,000,000 copies/mL
	Aptima HIV-1 RNA qualitative assay	TMA HPA	Qualitative	100 copies/mL
Qiagen	Artus CMV RGQ MDX kit	PCR	Quantitative	159−7.94 × 10,000,000 IU/mL

bDNA, Branched DNA signal amplification; *CMV*, cytomegalovirus; *CAP/CTM*, COBAS AmpliPrep/COBAS TaqMan; *HBV*, hepatitis B virus; *HCV*, hepatitis C virus; *HPA*, hybridization protection assay; *IU*, international unit; *LOD*, lower limit of detection; *PCR*, polymerase chain reaction; *RT-PCR*, reverse-transcriptase PCR; *TMA*, transcription-mediated amplification.

genotype worldwide and in Western countries is HCV genotype 1. Genotype 2 is found in Mediterranean countries and the Far East, genotype 3 in Europe, genotype 4 in the Middle East, genotype 5 in South Africa, and genotype 6 in Asia. HCV genotype information is an essential part of treatment strategies, especially with the newly available antiviral drugs [47].

HCV is a blood-borne pathogen, transmitted primarily by contact with infected blood, tissues, organs, and needles. Infection may also be transmitted through sexual intercourse and perinatally during child delivery. Acute HCV hepatitis may be asymptomatic or may produce nonspecific symptoms, such as malaise, nausea, dark urine, and jaundice. The incubation period varies between 0 and 6 months, and infection may resolve without any treatment. Chronic HCV infection develops in 50%−85% of acute hepatitis cases, and, without management, can lead to hepatic fibrosis, cirrhosis, end-stage liver diseases, and HCC. An estimated 170 million individuals are chronically infected with HCV worldwide with 3.2 million living with the disease in the United States. HCV is the most common reason for liver transplantation and the most common cause of death from liver disease [48].

Laboratory testing

Laboratory methods for the diagnosis of HCV infections include serology to detect total HCV antibodies and

molecular methods to detect HCV RNA [49]. Detection of HCV antibodies is useful for both screening and diagnosis, but serology does not distinguish between acute, chronic, or resolved infections. Current immunoassays represent the third generation for anti-HCV antibodies detection, and many are based on the detection of three-to-four recombinant HCV NS proteins, which increases the specificity of the assay over previous generations that used viral lysates to capture antibodies. One lateral flow immunoassay (OraQuick HCV, OraSure Technologies, Inc.) is available for point-of-care testing and is CLIA waived for testing on finger stick or venipuncture whole blood with a rapid TAT. The sensitivity and specificity of this rapid test are comparable to those of EIA assays. Screening immunoassays have high sensitivity and specificity. However, in low prevalence populations, specificity may only be moderate and, as such, confirmation of positive results is required (Fig. 54.1). Confirmation is done using an HCV RNA NAT assay (Table 54.7).

In addition to their utility for confirmation, NAT assays are used for monitoring treatment response (quantitative HCV NAT) and as prognostic assays to predict response to treatment (genotyping assays). Quantitative HCV NAT platforms target the NS regions of HCV and have linear ranges between 12 and 10^9 IU/mL. The lower limit of quantification is an important feature of these NATs, as definition of cure or sustained virologic response relies on a laboratory report of "undetectable" HCV RNA at the end of treatment and 6 months after the end of treatment. Current recommendation is that an appropriate HCV quantitative assay should have an LLOQ of at least 25 IU/mL with an LOD of 15 IU/mL [50]. Genotyping assays, which target the 5′ UTR of the HCV genome, are important to predict response to treatment as well as in therapy tailoring and for the determination of testing intervals. In addition to genotyping assays, HCV drug-resistance tests based on sequencing of the NS proteins NS3 and NS4A are important to optimize HCV therapy. Mutations or polymorphisms that confer resistance to direct-acting antivirals (DAAs) may be identified in treatment-naïve patients, in which case these mutations are referred to as baseline resistance-associated substitutions (RASs), or in patients treated with DAAs, which are referred to as treatment-emergent (or treatment-selected) RASs [51]. A combination of HCV genotype information and the presence or absence of specific RASs and other host factors (e.g., treatment-naïve patients or patients with cirrhosis) dictates the regimen of DAAs that will be used for each patient. For example, the treatment regimen for a treatment naïve genotype 1a patient without cirrhosis includes four recommended options (varying in length from 8 to 12 weeks) with four alternative DAA combinations (varying in duration from 12 to 16 weeks) [51]. The complexity of HCV treatment regimens underscores the importance of molecular tests in the management of HCV infections.

Previous CDC HCV screening recommendations to prevent HCV transmission were focused on screening of blood/organs donors and high-risk populations (i.e., injection drug users). In 2012 and 2013, the CDC released HCV screening guidelines that recommended a one-time testing, regardless of risk, for persons born between 1945 and 1965 presenting at a healthcare facility. This recommendation stemmed from surveillance data showing high

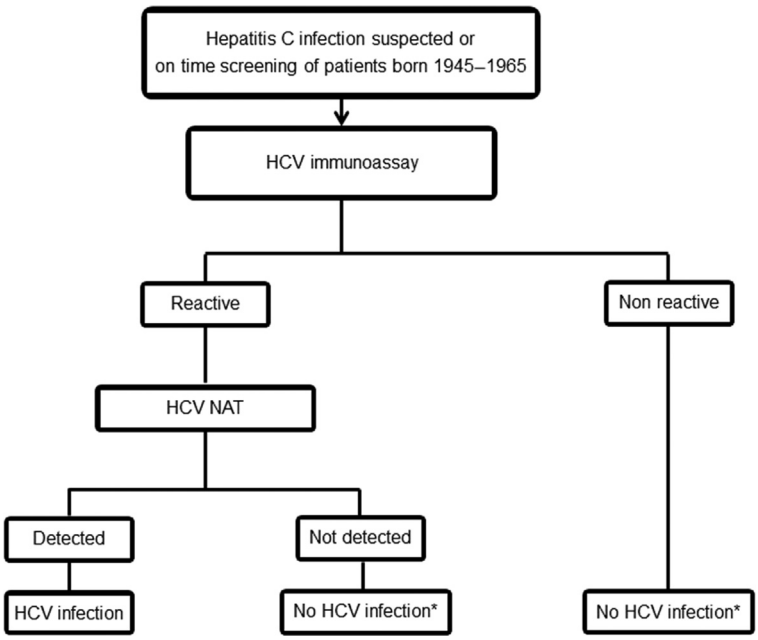


FIGURE 54.1 CDC recommended hepatitis C testing algorithm. *Additional testing should be done as appropriate.

prevalence of HCV in that age group, five times that of other adults' age groups. Increased testing should result in earlier identification of infected individuals at increased risk for complications of chronic HCV infections [48].

Hepatitis D

Hepatitis D virus (HDV) is an incomplete, single-stranded, circular RNA virus that requires the presence of HBV virus to replicate [52]. It is the only member of the *Delatviridae* family. Transmission of HDV occurs percutaneously or from contact with infected mucosal surfaces. The virus may be transmitted concomitantly with HBV virus (coinfection) or in patients with chronic HBV infection (superinfection). HDV may cause acute infections, which can be either asymptomatic or present with classic symptoms of hepatitis. Acute HDV infection is diagnosed based on symptoms and detection of HDV RNA (by reverse transcriptase PCR) or HDV antigen or anti-HDV IgM (by EIA/ELISA). In chronic HDV, total anti-HDV may be detected for several years along with HDV RNA or HDV antigen. Testing for HDV is not routinely performed in clinical laboratories and is often referred to specialized laboratories.

Hepatitis E

Hepatitis E virus (HEV) is a single-stranded RNA virus belonging to the *Herpesviridae* family. Four genotypes (HEV 1−4) have been identified, with HEV-1 and 2 prevalent in developing countries and HEV-3 and 4 more commonly found in developed countries. Like HAV, HEV is a nonenveloped virus, and, although it can survive in the environment for several months, it is more susceptible to heat and disinfectants than HAV. HEV is transmitted through the fecal−oral route, and the clinical presentation ranges from subclinical, asymptomatic infection to severe, fulminant hepatitis. Unlike HAV, HEV chronic infections have been reported, mostly in immunocompromised patients, such as SOT recipients. HEV has worldwide distribution, but it is highly endemic (i.e., >25% seroprevalence) in several countries in Africa, the Middle East, South-East Asia, and Central America [53].

Laboratory testing

In regions of low endemicity, like the United States, HEV should be considered only if travel history to a region of high endemicity is recorded or no other viral causes of hepatitis have been identified. The incubation period for HEV ranges from 3 to 8 weeks with anti-HEV IgM detectable within the first 4 weeks and falling to undetectable levels between 6 and 12 months. Anti-HEV IgG starts appearing a few days after the anti-HEV IgM, and, unlike anti-HAV IgG, they may become undetectable between 1 and 5 years following primary exposure.

There are no FDA-approved assays for the detection of HEV antibodies or HEV RNA. Acute HEV infection is diagnosed based on detection of anti-HEV IgM only or the detection of HEV RNA in serum and stool samples, which are detectable early after exposure and becomes undetectable approximately 3 and 5 weeks, respectively, after onset of symptoms. Detection of anti-HEV IgG alone reflects past infection. Commercial assays are available to detect HEV antibodies in serum samples (EIA/ELISA, rapid immunochromatographic, and point-of-care tests) and HEV RNA in serum and stool samples (reverse-transcriptase PCR). Although most of these assays target open reading frames 2 and 3 of the HEV, the use of different antigen preparations (recombinant vs synthetic antigens) has resulted in increased variability in the performance characteristics of serological HEV assays. In one recent study comparing five commercial assays, sensitivity of anti-HEV IgM and IgG ranged from 24% to 72% and 42% to 96%, respectively, with specificity of >96% for both IgM and IgG. HEV RNA was detected in several patients who were negative by serology, which prompted the authors of the study to recommend the use of HEV RNA assay to supplement diagnosis of acute HEV hepatitis [54].

Human immunodeficiency virus

Human immunodeficiency virus (HIV) is an enveloped, positive-sense, single-stranded RNA virus from the *Retroviridae* family, in the *Lentivirus* genus. The HIV genome is segmented, 9.7-kb long, with 9−10 genes coding several proteins including structural proteins: capsid protein (p24) encoded by the *gag* core gene; viral enzymes: protease (p12), reverse transcriptase (p66/p51), and integrase encoded by the *pol* gene; envelope proteins (gp 120 and gp 41) encoded by the *env* gene; and several regulatory proteins. Two species of HIV have been identified. HIV-1 is the most common virus worldwide and is divided into four groups (M, N, O, and P), several subtypes (A−K), and multiple circulating recombinant forms (CRFs). HIV-1 subtype B is the most common species found in North America and Europe. HIV-2 is found primarily in West Africa and is divided into two major groups (A and B) and a few subtypes and CRFs [55].

Transmission of HIV occurs through sexual contact, percutaneously, through blood transfusions with pathogen-containing blood, organ transplantation, or perinatally during vaginal delivery. Transmission through transfusion or organ transplantation occurs less frequently due to stringent screening of donors. Similarly, transmission from mother to newborn infant is less common in developed countries where women have better access to prenatal care and better

management of infants born to seropositive mothers, including prophylaxis or empiric HIV therapy (e.g., zidovudine, lamivudine, or nevirapine for high-risk transmission or zidovudine for low-risk transmission) [56]. Following infection, HIV binds primarily to CD4 T-lymphocytes present on the mucosal surface. Binding occurs between HIV viral envelope glycoproteins (e.g., gp 120) and the CXCR4 receptors or CCR5 chemokine receptors found on CD4 T-lymphocytes or macrophages, respectively. As is characteristic for retroviruses, the RNA genome is first converted by the viral reverse-transcriptase enzyme to a complementary DNA copy, which becomes integrated in the host cell genome, forming the HIV provirus. Furthermore, viral replication occurs using both host and viral replication enzymes, forming new virions that are propagated to other cells [57].

There are three clinical stages of HIV infection (Fig. 54.2). The first stage, acute or primary HIV, is characterized by flu-like symptoms of malaise and fever. This stage is highly infectious and lasts several weeks following viral acquisition. Recovery from acute HIV is followed by establishment of the infection, which is associated with mild or no symptoms. This stage can last for several years and, if not treated, can proceed to the third and last stage, acquired immunodeficiency syndrome (AIDS), where the CD4 T-cell count is low (<200 cells/mm^3) and opportunistic infections (e.g., pneumocystis) are common. Globally, the number of people living with HIV was reported at 36.9 million in 2017 with 1.8 million new infections and 940,000 AIDS-related deaths; many HIV-positive individuals are coinfected with HBV, HCV, or tuberculosis [58]. In the United States, approximately 40,000 cases of new HIV infections occurred in 2016, a decline of 5% from 2011 to 2015 [59]. Cases of AIDS have decreased steadily since the AIDS epidemic of the late 1980s, due in part to a combination of increased access to care, improved diagnostic tests, and efficacious combinatorial antiretroviral treatment (ART) regimens.

Laboratory testing

Diagnostically, the period of time immediately following infection is termed the eclipse period, since no diagnostic markers (antigens, RNA, or antibodies) can be detected during that time [60]. In weeks 1–4 after infection, HIV RNA concentrations increase drastically with levels reaching several millions copies per milliliter of plasma, a concentration that remains high for several months, and decreases to steady level for the duration of the asymptomatic phase of infection. Retrospective studies have shown that, in untreated hosts, the asymptomatic, chronic phase of HIV infection can last approximately 10 years, with some hosts (elite controllers) remaining asymptomatic for >20 years. This state is due in part to factors

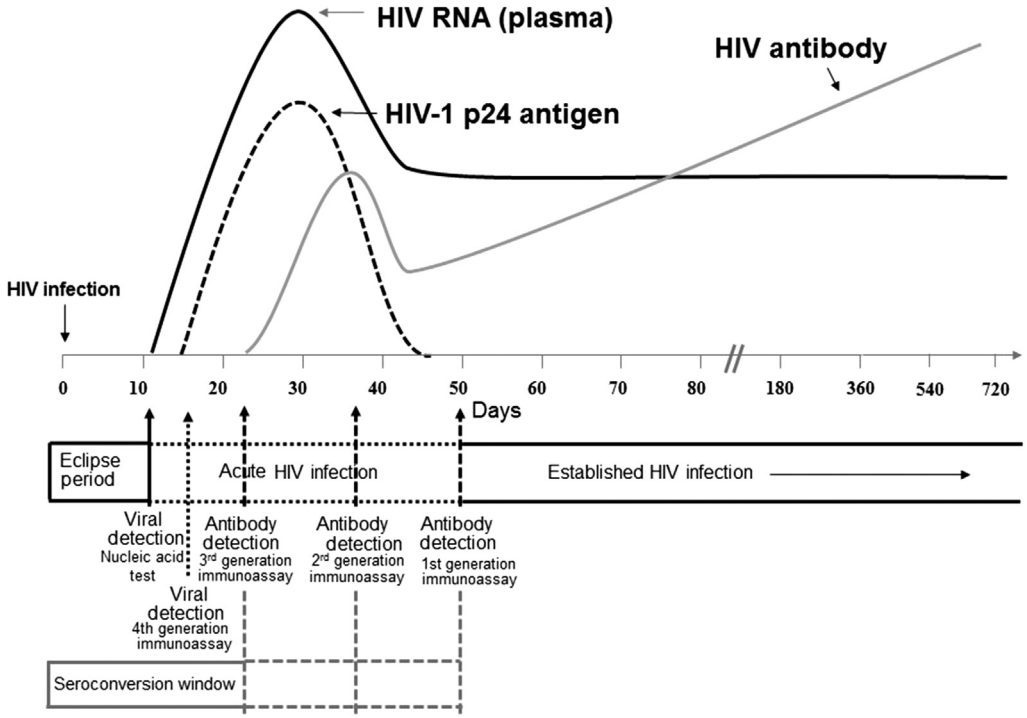


FIGURE 54.2 Stages of human immunodeficiency virus infections and corresponding markers. *Reproduced from the Centers for Disease Control and Prevention and Association of Public Health Laboratories, Laboratory testing for the diagnosis of HIV infection: updated recommendations. <http://stacks.cdc.gov/view/cdc/23447> (Published June 27, 2014; accessed 21.10.19).*

related to the virus itself (e.g., mutations) but is primarily a result of a balance between the host immune functions (e.g., T-cell diversity and mucosal tissues) and viral replication [61]. The p24 antigen will peak after 4 weeks and become undetectable by the end of the acute stage postviral acquisition. HIV antibodies start rising approximately a month after primary infection and steadily increase during the acute and asymptomatic phases, reaching peak levels during AIDS (Fig. 54.2).

Diagnosis of HIV infections and AIDS has improved significantly from the original tests, with improved sensitivity for both serological and NAT assays. Immunoassays are used for screening and diagnosis of HIV infections and can provide information of acute versus established infection. Traditionally, HIV immunoassays have been described in terms of assay generation, but current efforts aim at changing the nomenclature to focus on the target analytes being tested (e.g., IgG sensitive tests and p24 sensitive tests) [62]. The first-generation HIV immunoassays were designed using viral culture cell lysates and could only detect HIV-1 IgG antibodies (IgG sensitive tests). Those assays were eventually replaced with the second-generation assays that detected antibodies to both HIV-1 and HIV-2, but still, only IgG antibodies. The third-generation assays, which are still in use, provided the added advantage of detecting both IgM and IgG antibodies (IgM/IgG sensitive tests) to HIV-1 and HIV-2. These assays have improved specificity compared with the second-generation assays based on their use of synthetic or recombinant antigens. The fourth-generation immunoassays (p24/IgM/IgG sensitive assays) are designed to detect both the p24 antigen and the HIV-1/2 IgM and IgG antibodies. These assays utilize a combination of synthetic or recombinant antigens and monoclonal antibodies and allow for diagnosis of acute HIV to be shortened by up to 5 days (via p24 antigen detection), resulting in detection as early as 3 weeks postinfection (Fig. 54.2). Given that HIV viral loads are highest during the acute HIV phase, rapid and sensitive methods that identify infected individuals are key to limit transmission of the virus. An FDA-cleared fifth-generation HIV immunoassay is also available and differs from the fourth-generation assays in that it can detect and differentiate between the p24 antigen, the HIV-1 antibodies, and the HIV-2 antibodies in patient samples. A summary of the most common HIV immunoassays and their methodologies (EIA or CLIA) is listed in Table 54.6.

A few rapid HIV tests, including one fourth-generation test, are available as FDA-waived point-of-care tests. These rapid tests are based on lateral flow immunoassays or immunochromatographic methods. One test, the Geenius HIV Ag/Ab test (Bio-Rad), has replaced the HIV Western blot (WB), a first-generation test, as the recommended supplemental test following a reactive screen to differentiate between nonacute HIV-1 and HIV-2 infections (Fig. 54.3). Several studies have shown that rapid HIV tests have significantly higher accuracy and sensitivity than WB and detect seroconversion 7–15 days

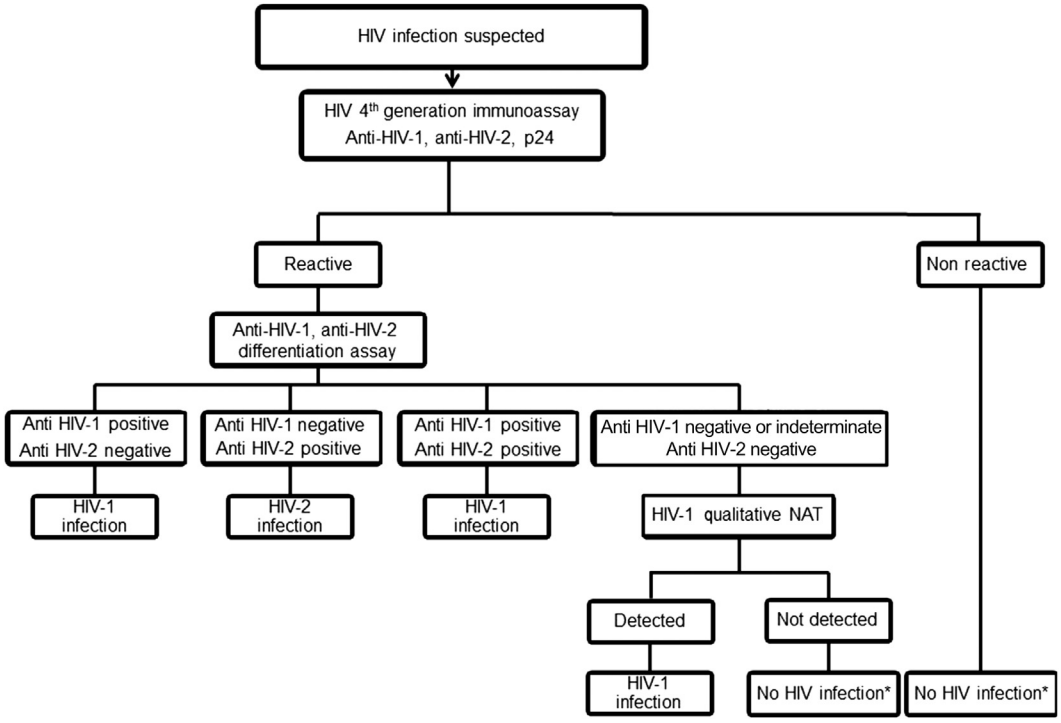


FIGURE 54.3 CDC and Association of Public Health Laboratories recommended testing algorithm for HIV infection. *Additional testing should be done as appropriate. *Available at <http://stacks.cdc.gov/view/cdc/50872> (Updated January 2018; accessed 21.10.2019).*

earlier than WB. However, current recommendations for the use of rapid, fourth-generation HIV tests vary depending on where testing is performed. While instrumented, the fourth-generation HIV tests are the preferred screening tests; if performed in a clinical laboratory, the point-of-care-based rapid fourth-generation HIV test may be used as the first step of the algorithm. However, if the rapid test is performed in a CLIA-waived setting and the result is positive, confirmation should be performed in a laboratory, starting with screening using an instrumented, fourth-generation HIV test [60,63,64].

HIV NAT assays serve multiple roles, including the following: (1) diagnosis of acute HIV infections (qualitative NAT); (2) monitoring of antiviral treatment response (quantitative NAT); and (3) detection of antiviral resistance (genotyping assays) [65]. Only one qualitative NAT assay, the Aptima HIV-1 RNA, is FDA-cleared for use in HIV diagnosis and as a supplemental assay to confirm a reactive fourth-generation antigen/antibody screen result. HIV-1 RNA has improved sensitivity when compared with WB for confirming HIV-1 acute infections, and decreases the eclipse period. HIV viral monitoring is used to establish a baseline HIV viral load and monitor successful response to ART. Several FDA-cleared quantitative HIV NAT assays are commercially available and all target HIV-1 viral loads. Such assays are generally optimized for group M subtypes (Table 54.7). Virologic blip (isolated detectable HIV RNA level that is followed by a return to virologic suppression) and low-level viremia (detectable HIV RNA of $<$200 copies/mL) may be early indicators of virologic failure (the inability to achieve or maintain suppression of viral replication to an HIV RNA level of $<$200 copies/mL) [66]. The LOD and the lower limit of quantification of these assays have continued to decrease, and most are as sensitive as the qualitative HIV-1 RNA tests, with values ranging from 20 to 175 copies/mL (Table 54.7). NAT assays are also useful for diagnosis of neonates and infants born to HIV-1 positive mothers. Molecular testing, either for proviral HIV DNA or for HIV RNA, is the recommended diagnostic approach for infants born of seropositive mothers, and current guidelines recommend testing at birth (for high-risk cases) or at 14–21 days, 1–2 months, and 4–6 months to confirm infection.

HIV-1 drug resistance testing is conducted more commonly using genotyping assays, although phenotypic assays are still available. Molecular tests have a faster TAT (days instead of weeks) and are technically easier to perform than phenotypic assays, which require growth of the virus in cell culture. Genotyping assays analyze protease and RT gene sequences to identify mutations known to confer resistance to the main classes of antiretroviral drugs (protease inhibitors, nucleoside reverse-transcriptase inhibitors, and nonnucleoside reverse transcriptase inhibitors). The assays generally include three steps: nucleic acid extraction, reverse-transcription PCR, and cycle sequencing of amplicons generated by PCR. Due to the heterogeneity of the HIV genome, most genotyping assays include several sets of sequencing primers for complete coverage of the protease and RT genomic sequences. Analysis of sequenced viruses is usually performed automatically by instrument software, and a report is generated that includes a list of mutations identified and corresponding drug resistance patterns. Genotyping assays are performed at baseline, prior to initiation of ART, and to guide alternative ART selection in cases of therapeutic failure. These assays are often performed in reference laboratories and often require a minimum viral load (e.g., 500 copies/mL) for analysis. Newer methodologies, including whole-genome sequencing, have increased sensitivity for detection of small populations of HIV-1 variants and may eventually circumvent the need for a minimum viral load [67].

HIV testing guidelines have changed drastically since the first CDC recommendations were published in 1989. The most current guidelines from the CDC and the Association of Public Health Laboratories recommend that initial testing for HIV diagnosis be performed using a laboratory-based fourth-generation immunoassay. A reactive screen should be followed by testing with a manual second-generation antibody test (Bio-Rad Geenius) to differentiate between HIV-1 and HIV-2. If the antibody test is negative for both HIV-1 and HIV-2, additional testing using an HIV-1 NAT should be performed (Fig. 54.3).

In addition to tests that target the HIV virus or antibody response, measurement of CD4T lymphocytes levels is an integral part of diagnosis and monitoring of HIV-infected patients. CD4 counts reflect the patient's immune recovery and should be measured in all patients at entry into care and at various intervals (every 3–6 months) during care depending on response to ART. In general, more frequent testing is recommended if CD4 counts are $<$200 cells/mL, and testing is optional if CD4 counts are $>$500 cells/mL and the patients is virologically suppressed (i.e., viral load is undetectable).

Conclusion

Herpesviruses and blood-borne viruses, including hepatitis and HIV viruses, have a significant impact on human health, causing a range of clinical syndromes from self-limited infections to chronic diseases. Screening, diagnosis, and monitoring of these infections rely heavily on both serological and molecular methods. As the performance characteristics of these methods continues to improve and novel treatments become available, laboratories will continue to be tasked to remain up to date on best methods and algorithms for rapid and accurate diagnosis and management of patients infected with these viruses.

References

[1] C. Jones, Herpes simplex virus type 1 and bovine herpesvirus 1 latency, Clin. Microbiol. Rev. 16 (2003) 79–95.

[2] C. Johnston, L. Corey, Current concepts for genital herpes simplex virus infection: diagnostics and pathogenesis of genital tract shedding, Clin. Microbiol. Rev. 29 (2016) 149–161.

[3] K.J. Looker, A.S. Magaret, M.T. May, K.M.E. Turner, P. Vickerman, L.M. Newman, et al., First estimates of the global and regional incidence of neonatal herpes infection, Lancet Glob. Health 5 (2017) e300–e309.

[4] M.A. Albrecht, Treatment of herpes zoster in the immunocompetent host. <https://www.uptodate.com> (accessed 21.10.19).

[5] C. Di Pentima, Viral meningitis: management, prognosis, and prevention in children. <https://www.uptodate.com> (accessed 21.10.19).

[6] G.J. Demmler-Harrison, Neonatal herpes simplex virus infection: management and prevention. <https://www.uptodate.com> (accessed 21.10.19).

[7] S. Ratnam, A. Severini, G. Zahariadis, M. Petric, B. Romanowski, The diagnosis of genital herpes - beyond culture: an evidence-based guide for the utilization of polymerase chain reaction and herpes simplex virus type-specific serology, Can. J. Infect. Dis. Med. Microbiol. 18 (2007) 233–240.

[8] A. Wald, R. Ashley-Morrow, Serological testing for herpes simplex virus (HSV)-1 and HSV-2 infection, Clin. Infect. Dis. 35 (2002) S173–S182.

[9] U.D. Allen, J.L. Robinson, I.D. Canadian Paediatric Society, C. Immunization, Prevention and management of neonatal herpes simplex virus infections, Paediatr. Child. Health 19 (2014) 201–212.

[10] M.B. Wilck, R.A. Zuckerman, Practice ASTIDCo. Herpes simplex virus in solid organ transplantation, Am. J. Transplant. 13 (Suppl 4) (2013) 121–127.

[11] N.W. Anderson, B.W. Buchan, N.A. Ledeboer, Light microscopy, culture, molecular, and serologic methods for detection of herpes simplex virus, J. Clin. Microbiol. 52 (2014) 2–8.

[12] D.W. Kimberlin, Neonatal herpes simplex infection, Clin. Microbiol. Rev. 17 (2004) 1–13.

[13] M.J. Binnicker, M.J. Espy, C.L. Irish, Rapid and direct detection of herpes simplex virus in cerebrospinal fluid by use of a commercial real-time pcr assay, J. Clin. Microbiol. 52 (2014) 4361–4362.

[14] R.M. Liesman, A.P. Strasburg, A.K. Heitman, E.S. Theel, R. Patel, M.J. Binnicker, Evaluation of a commercial multiplex molecular panel for diagnosis of infectious meningitis and encephalitis, J. Clin. Microbiol. (2018) 56.

[15] B.A. Forbes, Acquisition of cytomegalovirus infection: an update, Clin. Microbiol. Rev. 2 (1989) 204–216.

[16] V. Faure-Bardon, J.F. Magny, M. Parodi, et al., Sequelae of congenital cytomegalovirus following maternal primary infections are limited to those acquired in the first trimester of pregnancy, Clin. Infect. Dis. 69 (2019) 1526–1532.

[17] P. Ljungman, M. Boeckh, H.H. Hirsch, et al., Definitions of cytomegalovirus infection and disease in transplant patients for use in clinical trials, Clin. Infect. Dis. 64 (2017) 87–91.

[18] R.R. Razonable, A. Humar, Practice ASTIDCo. Cytomegalovirus in solid organ transplantation, Am. J. Transplant. 13 (Suppl. 4) (2013) 93–106.

[19] V. Erard, K.A. Guthrie, S. Seo, et al., Reduced mortality of cytomegalovirus pneumonia after hematopoietic cell transplantation due to antiviral therapy and changes in transplantation practices, Clin. Infect. Dis. 61 (2015) 31–39.

[20] P. Ljungman, M. Hakki, M. Boeckh, Cytomegalovirus in hematopoietic stem cell transplant recipients, Hematol. Oncol. Clin. North. Am. 25 (2011) 151–169.

[21] A.M. Caliendo, Overview of diagnostic tests for cytomegalovirus infection. <https://www.uptodate.com/contents/overview-of-diagnostic-tests-for-cytomegalovirus-infection> (accessed 21.10.19).

[22] M. Boeckh, G. Boivin, Quantitation of cytomegalovirus: methodologic aspects and clinical applications, Clin. Microbiol. Rev. 11 (1998) 533–554.

[23] M.J. Binnicker, M.E. Espy, Comparison of six real-time PCR assays for qualitative detection of cytomegalovirus in clinical specimens, J. Clin. Microbiol. 51 (2013) 3749–3752.

[24] H.H. Hirsch, I. Lautenschlager, B.A. Pinsky, L. Cardenoso, S. Aslam, B. Cobb, et al., An international multicenter performance analysis of cytomegalovirus load tests, Clin. Infect. Dis. 56 (2013) 367–373.

[25] R. Hoteit, P. Fermanian, F. Abbas, R. Abdel Khalek, R. Mahfouz, Comparison of the Artus RotorGene and COBAS AmpliPrep/COBAS TaqMan platforms for the detection of cytomegalovirus: experience of a tertiary care center, Genet. Test. Mol. Biomarkers 20 (2016) 167–169.

[26] H.P. Tsai, Y.Y. Tsai, I.T. Lin, P.H. Kuo, T.Y. Chen, K.C. Chang, et al., Comparison of two commercial automated nucleic acid extraction and integrated quantitation real-time PCR platforms for the detection of cytomegalovirus in plasma, PLoS One 11 (2016) e0160493.

[27] O.A. Odumade, K.A. Hogquist, H.H. Balfour Jr., Progress and problems in understanding and managing primary Epstein–Barr virus infections, Clin. Microbiol. Rev. 24 (2011) 193–209.

[28] A.W. Loren, D.L. Porter, E.A. Stadtmauer, D.E. Tsai, Post-transplant lymphoproliferative disorder: a review, Bone Marrow Transplant. 31 (2003) 145–155.

[29] L.S. Young, A.B. Rickinson, Epstein–Barr virus: 40 years on, Nat. Rev. Cancer 4 (2004) 757–768.

[30] M. De Paschale, P. Clerici, Serological diagnosis of Epstein–Barr virus infection: problems and solutions, World J. Virol. 1 (2012) 31–43.

[31] M.L. Gulley, W. Tang, Using Epstein–Barr viral load assays to diagnose, monitor, and prevent posttransplant lymphoproliferative disorder, Clin. Microbiol. Rev. 23 (2010) 350–366.

[32] R.T. Hayden, K.M. Hokanson, S.B. Pounds, M.J. Bankowski, S.W. Belzer, J. Carr, et al., Multicenter comparison of different real-time pcr assays for quantitative detection of Epstein–Barr virus, J. Clin. Microbiol. 46 (2008) 157–163.

[33] R.T. Hayden, Y. Sun, L. Tang, G.W. Procop, D.R. Hillyard, B.A. Pinsky, et al., Progress in quantitative viral load testing: variability and impact of the WHO Quantitative International Standards, J. Clin. Microbiol. 55 (2017) 423–430.

[34] M. Sathar, P. Soni, D. York, GB virus C/hepatitis G virus (GBV-C/HGV): still looking for a disease, Int. J. Exp. Pathol. 81 (2000) 305–322.

[35] Centers for Disease Control and Prevention, Prevention of hepatitis A through active or passive immunization: recommendations of the Advisory Committee on Immunization Practices (ACIP), MMWR Recomm. Rep. 55 (2006) 1–18.

[36] D. Daniels, S. Grytdal, A. Wasley, Centers for Disease C, Prevention, Surveillance for acute viral hepatitis - United States, 2007, MMWR Surveill. Summ 58 (2009) 1–27.
[37] E. Tennant, J.J. Post, Production of false-positive immunoglobulin M antibodies to hepatitis A virus in autoimmune events, J. Infect. Dis. 213 (2016) 324–325.
[38] F.J. Mahoney, Update on diagnosis, management, and prevention of hepatitis B virus infection, Clin. Microbiol. Rev. 12 (1999) 351–366.
[39] C.J. Chu, E.B. Keeffe, S.H. Han, R.P. Perrillo, A.D. Min, C. Soldevila-Pico, et al., Hepatitis B virus genotypes in the United States: results of a nationwide study, Gastroenterology 125 (2003) 444–451.
[40] Centers for Disease Control and Prevention, Viral hepatitis surveillance—United States. <https://www.cdc.gov/hepatitis/statistics/2017surveillance/index.htm>, 2017 (accessed 21.10.19).
[41] T. Gomes, M. Tadrous, M.M. Mamdani, J.M. Paterson, D.N. Juurlink, The burden of opioid-related mortality in the United States, JAMA Netw. Open. 1 (2018) e180217.
[42] G. Fattovich, F. Bortolotti, F. Donato, Natural history of chronic hepatitis B: special emphasis on disease progression and prognostic factors, J. Hepatol. 48 (2008) 335–352.
[43] M.F. Sorrell, E.A. Belongia, J. Costa, I.F. Gareen, J.L. Grem, J.M. Inadomi, et al., National institutes of health consensus development conference statement: management of hepatitis B, Ann. Intern. Med. 150 (2009) 104–110.
[44] E.E. Mast, H.S. Margolis, A.E. Fiore, E.W. Brink, S.T. Goldstein, S.A. Wang, et al., A comprehensive immunization strategy to eliminate transmission of hepatitis B virus infection in the United States: recommendations of the Advisory Committee on Immunization Practices (ACIP) part 1: immunization of infants, children, and adolescents, MMWR Recomm. Rep. 54 (2005) 1–31.
[45] A. Valsamakis, Molecular testing in the diagnosis and management of chronic hepatitis B, Clin. Microbiol. Rev. 20 (2007) 426–439.
[46] H.M. Lee, B.A. Banini, Updates on chronic HBV: current challenges and future goals, Curr. Treat. Options Gastroenterol. 17 (2019) 271–291.
[47] T.J. Liang, M.G. Ghany, Current and future therapies for hepatitis C virus infection, N. Engl. J. Med. 368 (2013) 1907–1917.
[48] B.D. Smith, R.L. Morgan, G.A. Beckett, Y. Falck-Ytter, D. Holtzman, C.G. Teo, et al., Recommendations for the identification of chronic hepatitis C virus infection among persons born during 1945–1965, MMWR Recomm. Rep. 61 (2012) 1–32.
[49] Centers for Disease Control and Prevention, Testing for HCV infection: an update of guidance for clinicians and laboratorians, MMWR Morb. Mortal. Wkly. Rep. 62 (2013) 362–365.
[50] A.-I.H.G. Panel, Hepatitis C guidance 2018 update: AASLD-IDSA recommendations for testing, managing, and treating hepatitis c virus infection, Clin. Infect. Dis. 67 (2018) 1477–1492.
[51] IDSA Aa, HCV guidance: recommendations for testing, managing, and treating hepatitis C. <https://www.hcvguidelines.org/sites/default/files/full-guidance-pdf/HCVGuidance_May_24_2018b.pdf> (accessed 21.10.18).
[52] C. Alves, C. Branco, C. Cunha, Hepatitis delta virus: a peculiar virus, Adv. Virol. 2013 (2013) 560105.
[53] N. Kamar, H.R. Dalton, F. Abravanel, J. Izopet, Hepatitis E virus infection, Clin. Microbiol. Rev. 27 (2014) 116–138.
[54] H. Norder, M. Karlsson, Å. Mellgren, J. Konar, E. Sandberg, A. Lasson, et al., Diagnostic performance of five assays for anti-HEV IgG and IgM in a large cohort study, J. Clin. Microbiol. (2015). In press.
[55] F.E. McCutchan, Global epidemiology of HIV, J. Med. Virol. 78 (Suppl 1) (2006) S7–S12.
[56] Panel on Treatment of HIV-Infected Pregnant Women and Prevention of Perinatal Transmission, Recommendations for the use of antiretroviral drugs in pregnant women with HIV infection and interventions to reduce perinatal HIV transmission in the United States. <https://aidsinfo.nih.gov/contentfiles/lvguidelines/perinatalgl.pdf> (accessed 21.10.19).
[57] M.S. Cohen, G.M. Shaw, A.J. McMichael, B.F. Haynes, Acute HIV-1 infection, N. Engl. J. Med. 364 (2011) 1943–1954.
[58] UNAIDS, 2017 global HIV statistics, 2018.
[59] Centers for Disease Control and Prevention, Estimated HIV incidence and prevalence in the United States, 2010–2015, in: HIV Surveillance Supplemental Report, vol. 23, 2018.
[60] Centers for Disease Control and Prevention, Laboratory testing for the diagnosis of HIV infection: updated recommendations. <https://doi.org/10.15620/cdc.23447> (accessed 21.10.19).
[61] A.A. Lackner, M.M. Lederman, B. Rodriguez, HIV pathogenesis: the host, Cold Spring Harb. Perspect. Med. 2 (2012) a007005.
[62] C.B. Hurt, J.A.E. Nelson, L.B. Hightow-Weidman, W.C. Miller, Selecting an HIV test: a narrative review for clinicians and researchers, Sex. Transm. Dis. 44 (2017) 739–746.
[63] Centers for Disease Control and Prevention, 2018 quick reference guide: recommended laboratory HIV testing algorithm for serum or plasma specimens. <https://stacks.cdc.gov/view/cdc/50872> (accessed 21.10.19).
[64] Centers for Disease Control and Prevention, Technical update: use of the determine HIV 1/2 Ag/Ab combo test with serum or plasma in the laboratory algorithm for HIV diagnosis. <https://stacks.cdc.gov/view/cdc/48472> (accessed 21.10.19).
[65] Y.W. Tang, C.Y. Ou, Past, present and future molecular diagnosis and characterization of human immunodeficiency virus infections, Emerg. Microbes Infect. 1 (2012) e19.
[66] Services DoHaH. Panel on antiretroviral guidelines for adults and adolescents. Guidelines for the use of antiretroviral agents in adults and adolescents living with HIV. Department of Health and Human Services. <http://aidsinfo.nih.gov/contentfiles/lvguidelines/AdultandAdolescentGL.pdf> (accessed 04.07.18).
[67] J. Weber, I. Volkova, M.K. Sahoo, P.L. Tzou, R.W. Shafer, B.A. Pinsky, Prospective evaluation of the vela diagnostics next-generation sequencing platform for HIV-1 genotypic resistance testing, J. Mol. Diagn. (2019).

Self-assessment questions

1. Which of the following properties is common to all herpesviruses?
   a. Responsible for causing cold sores
   b. Establish latency in specific cell types
   c. Diagnosed with high sensitivity using immunoassay
   d. Viral genome is a single-stranded RNA genome
   e. Viral genome is a single-stranded DNA genome
2. Which of the following assays has the highest sensitivity for diagnosis of HSV meningitis?
   a. IgM and IgG antibodies immunoassays on serum
   b. IgM and IgG antibodies immunoassays on CSF
   c. IgM immunoassays on CSF
   d. PCR assay on CSF
   e. PCR assay on serum
3. Which of the following markers is detectable in patients with a past EBV infection?
   a. VCA IgM
   b. VCA IgG
   c. VCA IgA
   d. EBNA-1 IgG
   e. EA IgG
4. Which of the following is a limitation of CMV antigenemia?
   a. High sensitivity in neutropenic patients
   b. Labor-intensive test
   c. Long turnaround time
   d. Low specificity
   e. High cost
5. Which of the following hepatitis viruses is transmitted through the fecal−oral route?
   a. HCV
   b. HBV
   c. HDV
   d. HEV
   e. HGV
6. Which of the following markers is detectable in an HBV vaccinated person?
   a. HBs antigen
   b. HBs antibodies
   c. Total HBc antibodies
   d. HBc IgM
   e. HBe antibodies
7. Which of the following tests is recommended for initial screening of HIV infection?
   a. HIV-1/2 fourth-generation immunoassay
   b. HIV-1/2 third-generation immunoassay
   c. HIV-1/2, p24 fourth-generation immunoassay
   d. HIV-1 PCR
   e. HIV-1/2 lateral flow immunoassay

Answers

1. b
2. d
3. d
4. b
5. d
6. b
7. c

Chapter 55

Clinical microbiology

Morgan A. Pence[1] and Rachael Liesman[2]

[1]*Clinical and Molecular Microbiology, Department of Pathology & Laboratory Medicine, Cook Children's Medical Center, Fort Worth, TX, United States,* [2]*Clinical Microbiology, Department of Pathology & Laboratory Medicine, The University of Kansas Health System, Kansas City, KS, United States*

Learning objectives

After reviewing this chapter, the reader will be able to:

- List key considerations for specimen collection for microbiology testing.
- Discuss the advantages and limitations of automation in the clinical microbiology laboratory.
- Describe the evolution of microorganism identification methods.
- Discuss the benefits and limitations of molecular microbiology point-of-care testing.
- Summarize currently available multiplex molecular microbiology testing options.

Specimen collection

Specimen collection in clinical microbiology is of utmost importance. The quality of the specimen determines the quality of the results. Proper specimen collection consists of (1) proper collection of the source; (2) proper container selection; and (3) proper transport conditions. Microbiology laboratory staff consists of technicians and technologists who are thoroughly trained on the appropriate processing and testing methodologies; however, even the most skilled microbiologists and the best laboratory practices cannot make up for a poor specimen. Although improperly collected specimens lead to unreliable results, they are received in the microbiology laboratory every day. Thus it is the job of the laboratory to convey information about proper collection techniques and containers to clinicians.

To communicate and highlight the importance of obtaining appropriate specimens for microbiological testing, the Infectious Diseases Society of America and the American Society for Microbiology (ASM) published a guidance document on utilization of the microbiology laboratory, which contains information on optimal test selection, optimal specimen collection approaches, and transport concerns [1]. In addition, the Guide to Specimen Management in Clinical Microbiology, now in its third edition, is available from ASM Press [2]. Both documents are critical resources for additional information regarding management of specimens in microbiology laboratories.

In general, aspirates, fluids, and tissue specimens are preferred over swabs for all microbiology culture testing due to their higher diagnostic yield. However, it is not always possible to obtain an aspirate, fluid, or tissue, so swabs are commonly accepted. Disinfection of the site (if applicable) must be carefully considered during specimen collection. For example, it is imperative to disinfect the skin prior to collection of blood cultures. Otherwise, skin flora may result in a false-positive culture, which can lead to an increase of $2923 to $5812 in hospital costs as well as exposure to unnecessary antimicrobials and increased length of stay [3]. Collection from the appropriate source is also an important consideration. If a nasopharyngeal (NP) swab is the preferred source for a test, it is important to collect a true NP swab, which is not an enjoyable experience for the patient, rather than a nasal swab. For example, *Bordetella pertussis* and many respiratory viruses are primarily found in the nasopharynx, so a properly collected NP swab is essential for detection and diagnosis. A swab of the nares or a mid-turbinate region may cause a false-negative result. Blood cultures and NP swabs are only such two examples to highlight the importance of appropriate source collection.

In comparison with the core laboratory, the microbiology laboratory receives a much wider variety of transport containers, which may include Tupperware containers, Ziploc bags, Mountain Dew bottles, etc. (all of which have been received in a clinical microbiology laboratory). Specimens received in such nonsterile containers are rejected, but the variety of acceptable sterile transport devices can still be overwhelming and collection containers are not standardized between the laboratories due to the extensive number of transport device manufacturers

Contemporary Practice in Clinical Chemistry. DOI: https://doi.org/10.1016/B978-0-12-815499-1.00055-7

and variations in devices among manufacturers. One laboratory may use one set of collection devices, while other laboratories have their own sets. A standalone hospital microbiology lab may receive as few as 10−20 different types of containers, consisting of various sterile cups and tubes, preservative tubes, capped syringes, swabs, etc., while the number of collection devices received at the centralized and reference laboratories may be much larger. The variability in specimen collection containers was a major barrier to implementation of total laboratory automation (TLA) for clinical microbiology workflows.

Specimens for anaerobic culture should be submitted under conditions that allow recovery of anaerobes. For example, anaerobic transport containers that contain a semisolid reducing gel may be used for specimens submitted for anaerobic culture. In addition, capped syringes with excess air removed may be used. Dry swabs are not appropriate for anaerobic culture. Anaerobic culture testing should never be performed without an accompanying aerobic culture, unless selective culture for *Clostridioides difficile* is requested; however, this is uncommon. Specimens acceptable for anaerobic culture include aspirates, tissues, and deep wounds. Superficial wounds and other sites with normal anaerobic flora are not acceptable for anaerobic culture.

One of the major shifts in clinical microbiology was the development of the flocked swab in transport media. Historically, tightly wound cotton, rayon, and dacron swabs predominated, with some placed into a liquid or gel transport media while others were not. Traditional swabs performed poorly, because the vast majority of the organisms in the specimen remained trapped in the swab and were not released when the swabs were used to inoculate solid media (Petri dishes). Another drawback of traditional swabs is that if multiple plates need to be inoculated, the majority of the bacteria that are released were released onto the first plate, resulting in inconsistent inoculation across culture media. Compared with traditional swabs, flocked swabs contain fibers that radiate outward from the shaft, which allow for increased release of organisms from the swab. Once the specimen is collected, flocked swabs are placed into transport tubes, which contain liquid transport media. While dry swabs cannot be used for anaerobic culture due to anaerobes drying out and dying, the addition of transport media allows the swabs to be used for anaerobic culture in addition to aerobic culture. The transport media also permits a longer transit time to the lab due to improved specimen stability (24−48 hours). In addition, this setup allows for the release of organisms into the transport media, which can be used for the inoculation of plates rather than inoculating with a swab, allowing for consistent inoculation across culture media. Not only did the development of the flocked swab improve the quality of cultures, it also helped pave the way for automation in microbiology, because it is easier to automate the transfer of transport media to inoculate plates as compared with using a dry swab for plate inoculation. Commercially available flocked swab options include the Copan ESwab and Puritan PurFlock Ultra and HydraFlock swabs.

For urine testing in the microbiology laboratory, urine preservative tubes have become more common and can preserve organisms for up to 24−48 hours during transport. Preservative tubes are more standardized than sterile cups, which are different sizes and have different lid-threading properties. Transport media have been developed for stool, facilitating downstream automation and molecular testing of stool specimens. The shift toward transport media has allowed for improved culture results and support of automation. Although specimen containers have become more amenable to automation, there will always be containers that are not accommodated on automated instrumentation. For additional information, refer to the following section on laboratory automation.

After collection, the specimen should be transported to the laboratory in a timely manner. For off-site laboratories, specimens should be transported to maintain specimen integrity and quality. Excessive transport times may negatively affect results due to either the death of fastidious organisms or overgrowth of nonfastidious organisms. Generally, it is recommended that specimens are received in the laboratory within 1−2 hours after collection unless specimens are in a transport or preservative media, which permits longer transit times. Alternatively, some specimen sources, but not all (e.g., cerebrospinal fluid), can be refrigerated to preserve their quality. Transport time is also important, because the sooner the specimen arrives in the lab, the sooner the culture is inoculated and placed into an incubator. Delays in transport lead to delays in culture results.

Once specimens are received in the laboratory, they are processed in a biosafety cabinet to prevent staff exposure to infectious aerosols, which may be generated during specimen processing, and to prevent contamination from the environment. Only one specimen is processed at a time in order to prevent cross-contamination. This is important to ensure the quality of reported results, both for traditional microbiology and for molecular microbiology. Specimens processed in other areas of the laboratory outside of a biosafety cabinet should not be used for microbiology culture or molecular microbiology testing due to the possibility of false-positive results due to cross-contamination. Shared specimens should be processed initially in the microbiology laboratory prior to other testing (such as Core laboratory testing) or a separate specimen should be obtained.

A Gram stain is performed on the majority of specimens, with the exception of urines, stool, screening

cultures [e.g., Group B *Streptococcus* and methicillin-resistant *Staphylococcus* (MRSA), etc.], and strep throat swabs. A Gram stain can be performed on direct specimens, such as respiratory specimens, fluids, and tissues and on swabs, such as wound swabs. However, Gram stains performed on specimens collected on swabs produce inferior results compared with those performed on direct specimens. The Gram stain provides the clinician with an early result while culture results are pending. In addition to the evaluation of bacteria, Gram stains are analyzed for the presence of white blood cells (WBCs) and squamous epithelial cells (SECs). The presence or absence of WBCs and SECs determines the quality of the specimen. The Gram stain plays a significant role in determining workup for respiratory and wound cultures. As an example, lower respiratory specimens with ≥ 25 SECs per low power field should be rejected and not cultured. The presence of SECs indicates that the specimen is from the oropharynx rather than the lower respiratory tract.

Rejection criteria are essential, and each microbiology laboratory should have a procedure that addresses specimen rejection in order to avoid providing inaccurate results. Examples of rejection criteria that are commonly used are:

- Specimens submitted in nonapproved containers should be rejected.
- Leaking specimens should be rejected.
- Anaerobic culture, if requested, should not be performed if the specimen is not submitted under anaerobic conditions.
- Specimens with excessive transport times (defined by the laboratory, will depend on source) should be rejected.
- Respiratory specimens with ≥ 25 SECs per low power field on Gram stain should be rejected.

Laboratory automation in clinical microbiology

Several components in the clinical microbiology laboratory have been automated during the last few decades. Automation advances include automated blood culture instruments, automated antimicrobial susceptibility testing platforms, automated nucleic acid extraction, and others. Although components of the laboratory were automated, the primary role of the microbiology laboratory—specimen inoculation and culture automation—was not automated until recently due to a number of factors. The wide variety of specimens and specimen collection devices was a key factor as to why the development of TLA in microbiology took so long to become a reality. In addition to the variety of containers, specimens are processed in a variety of manners based on the specimen source and the test ordered. Specimens may be vortexed, centrifuged, minced, or ground. Moreover, once inoculated, plates are incubated in different temperatures under various atmospheric conditions. For decades, it seemed impossible to automate microbiology due to these aforementioned challenges. However, there was increasing need and demand for automation due to a nationwide shortage of medical technologists and due to the consolidation of microbiology

FIGURE 55.1 Components of total laboratory automation in clinical microbiology.

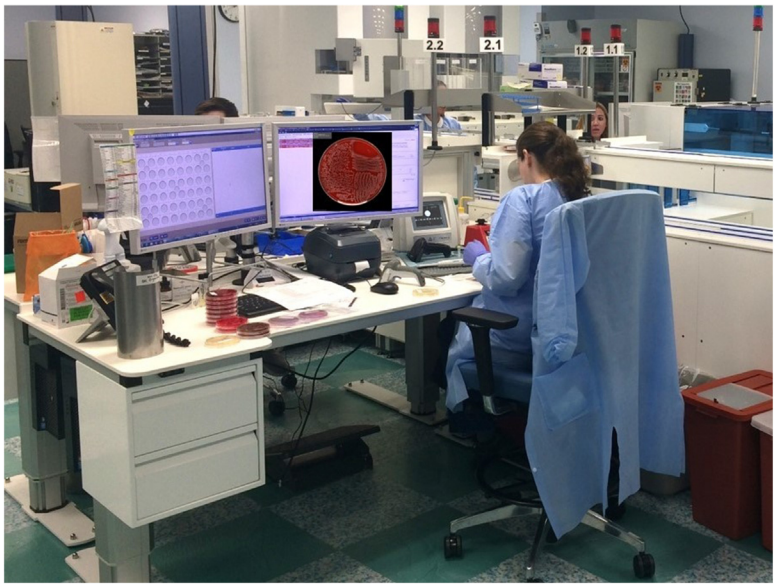


FIGURE 55.2 Total laboratory automation systems present high-resolution images on a computer monitor for technologists to read and interpret.

laboratories, which has resulted in increased workloads. In addition, a greater emphasis has been placed on quality and standardization in healthcare. These factors, along with recent technological advances, led to the rapid evolution of laboratory automation in clinical microbiology. An overview of laboratory automation in clinical laboratories is further described in Chapter 14, Laboratory automation, of this book.

Currently, there are two commercially available laboratory automation systems for clinical microbiology: Becton Dickinson (BD)'s Kiestra TLA and Copan Diagnostic's Walk Away Specimen Processor Laboratory automation system (WASPLab). Both systems automate all steps of the culture process, including specimen inoculation, incubation, and plate imaging (Fig. 55.1). The images are then presented on a computer monitor to technologists who read and interpret the cultures (Fig. 55.2). Both manufacturers are working to automate culture interpretation as well, but it is likely that there will always be a certain degree of human intervention required.

Components of automation

Inoculation unit

The processing components of the systems have been designed to accept multiple types of specimen containers and also allow for "offline" inoculation in case a nonstandard/nonprogrammed container is received. The systems vortex and centrifuge (available with the WASPLab system) specimens, uncap and recap specimen containers, and remove and replace lids on culture plates or broth-based media and inoculate specimens. After samples are added to plates, the systems streak plates according to programmed patterns that are preselected by the laboratory. In addition to processing and inoculation, both systems also prepare Gram stain slides for staining.

Automated track

Once inoculated, plates are transported via automated track lines to automated or "smart" incubators, which can be programed at various temperatures and atmospheric conditions. Plates may then be summoned from the incubators to workbenches via the automated track for further workup by the technologist. Systems may have a unidirectional or bidirectional track. Systems with a unidirectional track allow plates to be summoned to workbenches, assuming workbenches are downstream of the incubators, but plates must be placed in a canister and manually returned to the track upstream of the incubators in order to be returned to the incubator. Systems with a bidirectional track allow plates to be transported back to incubators directly from the workbenches once workup is completed.

Automated "smart" incubators

Prior to automation, plates were stacked in racks and placed in traditional incubators. Plates toward the top of the rack were more exposed to the proper incubation conditions (temperature and atmospheric conditions) compared with plates in the middle and bottom of each rack. In addition, incubator doors were opened and closed each time a plate was added or removed, leading to inconsistency in thermal conditions and other incubator parameters.

"Smart" incubators can automate all forms of aerobic incubation. Plates enter and exit incubators via the

automated track, so incubator doors remain closed throughout the process. This allows the incubators to maintain a constant temperature and environment. In addition, automated incubators contain individual shelves for each plate, which promotes faster, more consistent growth of colonies due to the consistent incubation conditions across all plates. Automated incubators can be programmed at a variety of temperatures and can maintain either an oxygen-rich or a carbon dioxide-enriched environment. Nonfastidious organisms, such as the *Enterobacterales*, grow well in an oxygen-rich environment, while a carbon dioxide environment promotes the growth of fastidious organisms. Moreover, carbon dioxide interferes with the performance of selective culture media. Selective media contain antibiotics and/or inhibitors to suppress the growth of certain organisms. An example is mannitol salt agar, which allows for the growth of *Staphylococcus aureus* while suppressing growth of other organisms. Such media should be incubated in an oxygen-rich environment. It is important to note that current systems do not offer an anaerobic incubator, so anaerobic culture plates must be incubated offline. Inside the incubators, a robotic arm facilitates the movement of the plates, by either placing them on a shelf or removing them for imaging or discarding. The imaging systems are a component either of the incubator or adjacent to the incubator, so plates are removed for a minimal amount of time when imaged.

Imaging

Automation systems are equipped with high-resolution cameras that capture images of plates at defined intervals, which are programmed by the laboratory. Imaging workflows vary based on culture type. Traditionally, technologists would hold plates up toward a light source and at various angles when viewing; this allowed easier detection of certain colony morphologies. Automated systems capture images using various lighting angles, including back lighting, front lighting, and side lighting to mimic manual viewing. Captured images are maintained on the system and can be viewed via remote access. In addition, the captured images allow technologists to go "back in time" when reading cultures. If growth on the second or third day differs from what was reported the first day, the technologist can review the previous days' images for comparison. The storage of image files also creates an opportunity for training of new technologists or to emphasize concepts. Moreover, stored images create an opportunity for quality assurance review.

Workstations

After the images are captured, technologists read plates by viewing the images on a computer rather than viewing the actual plates. Individual plate images can be enlarged to the size of the computer screen and further magnified. Because technologists are never exposed to a plate before viewing an image, imaging could potentially prevent exposure to a select agent (i.e., *Bacillus anthracis*, *Brucella* species, and others) or other highly infectious organisms, such as *Neisseria meningitidis*, based on suspicious growth characteristics. When a plate is viewed, the technologist can virtually mark specific colony morphologies for additional workup processes, which may include identification, antimicrobial susceptibility testing, and subculturing. Additional workup can be completed at the same time by summoning the plate to the workbench or, alternatively, the culture can be added to a queue and batched for workup after viewing is completed.

With the onset of laboratory automation, many microbiology laboratories have adopted a reader/workup workflow where culture images are read by one technologist and annotated for workup and positive cultures are followed up by a second technologist. This differs from the traditional microbiology laboratory, where one technologist would perform all reading and workup for a given culture. The benefit of the reader/workup workflow is that two technologists work together on the same culture, which may lead to improved consistency. This workflow also allows for off-site reading, so technologists could read cultures while in another area of the lab or offsite.

Commercially available systems

The two commercially available laboratory automation systems for microbiology are the Copan WASPLab and BD Kiestra TLA, as mentioned above. Both systems are modular, scalable, and offer automated selection and labeling of media, inoculation, incubation, and plate imaging. The systems interface with the laboratory information system (LIS), and when a specimen barcode is scanned, the LIS is queried and communicates the type of culture that was ordered. The system then selects the corresponding media (previously programmed by the laboratory) based on the type of culture ordered, applies a specimen label to each piece of media and inoculates the specimen.

The WASPLab inoculation unit, the WASP DT, uses 1-, 10-, and 30-μL reusable inoculating loops. Loops are sterilized between the pieces of media or between the specimens, depending on which program the lab selects. The Kiestra inoculation unit, InoculA, uses a pipettor rather than loops for inoculation. The InoculA pipettor can pipette 10−250 μL, and once a specimen is plated, the InoculA uses a single-use sterile magnetic bead to streak each plate. In both cases, specimens must be relatively nonviscous and homogenous for automated inoculation. The Kiestra system has the added consumable cost due to

the single-use nature of the beads, but produces more isolated colonies and more accurate colony counts than the WASPLab [4,5]. Both systems allow for manual inoculation if tissues, viscous specimens, or unusual containers are received. The original WASPLab requires offline manual inoculation and streaking, and plates can subsequently be placed onto the automated track for transport to the incubators. Copan recently developed a robot, called the Collaborative Robot, which can be purchased as a separate unit to assist with manual cultures. Technologists manually inoculate plates, and the Collaborative Robot streaks and places them on the WASPLab automated track. The Kiestra inoculation unit has a biosafety cabinet, in which system-labeled plates are presented for the technologist to inoculate manually. This is referred to as a semiautomated mode and is equivalent to Copan's Collaborative Robot. Once inoculated, the Kiestra system streaks the plates and processes them in the same manner as fully automated specimens. Both systems can create Gram stain slides, although staining and reading of smears are still performed manually.

While both systems have an automated track system, the Kiestra track is bidirectional, while the WASPLab track is unidirectional. The bidirectional track associated with the Kiestra TLA permits plates to be summoned to workbenches and subsequently returned to incubators directly from workbenches. Technologists at WASPLab workbenches can summon plates from incubators, assuming the benches are downstream of the incubators, but must manually place plates on the track upstream of the incubators for return. Both systems have the option of having workbenches that are separate from the automated track. This setup requires manual placement and retrieval of plates, but may be beneficial to labs with a smaller footprint.

Benefits of lab automation and future directions

Improved turnaround times

In traditional microbiology, there is significant variability in the inoculation efficiency between the microbiology staff members. The consistent inoculation of laboratory automation systems results in a higher number of isolated colonies, which are required for identification and susceptibility testing [5,6]. In traditional microbiology, organisms often have to be subcultured to obtain isolated colonies for additional testing, which increases turnaround times. Additionally, staff inoculate several cultures before transporting plates to the incubator, which increases the time the plates are incubated in a suboptimal environment. With automated systems, inoculated plates are transported along the automated track directly to incubators, which provide ideal growth conditions. Moreover, with automation technologists read and interpret images, rather than actual plates, so plates remain in optimal incubation conditions, rather than sitting outside of the incubator for hours at a time.

Due to a combination of the factors listed above, several studies have shown a decrease in the turnaround times of 2−30 hours across multiple culture types on WASPLab and Kiestra [7−11]. However, automation alone is insufficient to improve turnaround times [12]. In addition to automation, workflow changes are required to achieve efficiencies. The majority of laboratories that have implemented automation have also implemented plate reading on the second and/or third shift rather than limiting plate reading to first shift as has traditionally been the case in clinical microbiology laboratories. This allows cultures to be read, as they are ready rather than batching cultures for a single shift.

Automated culture reading

Artificial intelligence (AI) software, PhenoMATRIX, is available for WASPLab. PhenoMATRIX performs automated reading, interpretation, and segregation of cultures into growth versus no growth categories [13,14]. Results are grouped and presented to the technologist for final interpretation and resulting. The software is available for multiple culture types, but each version must be purchased separately. Multicenter studies using chromogenic agar for detection of MRSA and vancomycin-resistant *Enterococcus* (VRE) surveillance cultures demonstrated 100% sensitivity and 89.5%−96% specificity [13,14]. AI software is not currently available for Kiestra but is in development. A proof-of-concept study demonstrated that the Kiestra OPTIS software is able to provide reliable quantitative urine colony counts [15]. In addition to current AI capabilities, there are plans to develop AI that can classify or preliminarily identify organisms based on their morphologies.

TLA for clinical microbiology is still in its early stages. As automation technologies continue to improve, the clinical microbiology laboratory will continue to optimize its operation and workflow. This will result in increased efficiencies that address the shortage of medical technologists, provide a consistent level of care in a healthcare environment that is increasingly focused on quality care, and will continue to improve turnaround times.

Methods for identification of microorganisms

Once an organism is recovered in culture, whether via traditional microbiology or via an automated system, it may

be subjected to a downstream identification method. Such methods include biochemical testing, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS), or DNA sequencing. DNA sequencing and MALDI-TOF MS can also directly identify a pathogen from a primary specimen.

DNA sequencing

DNA sequencing is the gold standard for microorganism identification. The 16S ribosomal RNA (rRNA) gene is the most common sequencing target for bacteria and is 1542 base pairs (bp) in length. Sequencing the first 500 bp often provides enough differentiation for identification purposes, but certain organism groups require sequencing of the full-length 16S rRNA gene for differentiation, while other genera are homologous across the entire gene and require sequencing of additional genes for differentiation. The additional gene(s) required depends on the genus, and examples include *rpoB, recA, tuf, gyrA gyrB*, and *cpn60.* Many clinical laboratories that perform 16S sequencing only sequence a portion of the 16S rRNA gene, while few sequence the entire 1542 bp. Most clinical laboratories do not perform sequencing of the supplemental genes. Once a sequence is obtained, it is compared with a public or private database for organism identification. Database selection is critical, as some databases are curated and routinely updated and others are static. For more information on selecting an appropriate database, refer to the Clinical and Laboratory Standards Institute (CLSI) MM18 document (Interpretive Criteria for Identification of Bacteria and Fungi by Targeted DNA Sequencing, second edition) [16].

Interpretive criteria are based on the aforementioned CLSI MM18 document [16]; $\geq 97\%$ homology is required for genus-level identification of bacteria, while $\geq 99\%$ homology is required for species-level identification. Less than 95% homology indicates an incomplete database or potentially a novel species. Although 16S rRNA sequencing is the most common, it does not provide sufficient differentiation for particular groups of bacteria, such as *Escherichia coli* and *Shigella* spp., which are essentially the same organism, and *Bordetella* spp. In such cases, sequencing of additional genes or biochemical reactions is required for complete identification.

For yeast and filamentous molds, the most common sequencing targets are the internal transcribed spacer regions (ITS), ITS1 and ITS2. ITS1 and ITS2 are variable regions located between the conserved rRNA genes. ITS2 alone is sufficient for discriminating multiple *Candida* spp. but not for other yeasts or molds. In addition to ITS1 and ITS2, the D1/D2 region of the 28S rRNA gene may be used. A consensus has not been reached in regard to cutoff values for genus- and species-level identification of fungi.

Although DNA sequencing is the gold standard, the methodology is technologically challenging, has a slow turnaround time, is not widely available, and is relatively expensive. Because of these drawbacks, it is not routinely used to identify microorganisms and is more commonly used as a backup or secondary method to other identification methods.

Biochemical/phenotypical methods

Biochemical reactions and phenotypic characteristics have historically been used to identify microorganisms. Initially, biochemical reactions were carried out and analyzed in separate reaction tubes. However, this was burdensome, and some of the reactions required a significant amount of time. Commercial vendors miniaturized biochemical testing, which allowed multiple biochemical reactions to be combined into a panel. Commercial biochemical panels are interpreted manually by the technologist or automatically by an instrument. Examples of manual panels include ThermoFisher's Remel RapID series and bioMerieux's API series (Fig. 55.3). Automated instruments include bioMerieux's VITEK 2 (Fig. 55.4A and B), BD's Phoenix, and Beckman Coulter's MicroScan. For panels that require manual interpretation, the technologist visually determines the reactions and enters the results into a web-based database. Both manual and automated systems compare the reactions to an organism database and provide an identification with a confidence score, which is reported as a percentage.

Biochemical methods are inexpensive and easy to perform, but they are generally slow and have variable performance. Correct genus- and species-level identification rates for aerobic Gram-negative rods are 83% and 75%, respectively [17]. While biochemical methods perform relatively well for commonly isolated aerobic bacteria, they are not reliable for inert, slowly growing or infrequently isolated bacteria, including anaerobes. Correct genus-level identification rates range from 71% to 87%

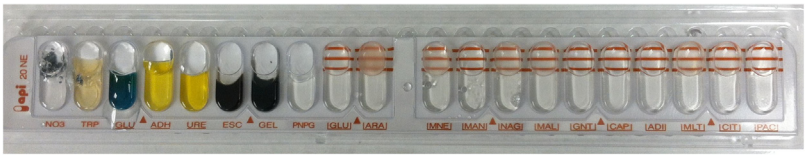


FIGURE 55.3 A bioMerieux API panel, an example of a miniaturized biochemical identification panel.

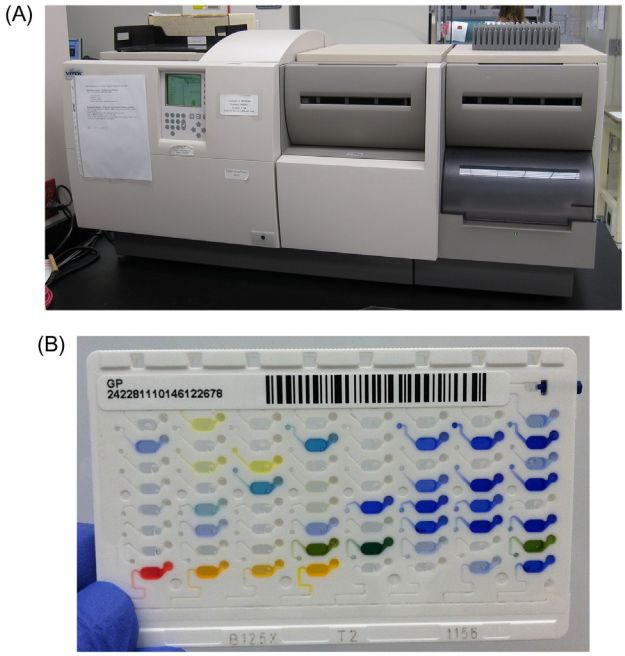


FIGURE 55.4 (A) VITEK 2 instrument and (B) VITEK 2 Gram-positive identification panel are shown. Each well on the identification panel represents a unique biochemical reaction.

for anaerobes, while correct species-level identification rates are a mere 50% to 60% [18–21]. Identification rates for infrequently isolated aerobic bacteria are even more dismal and have been reported to be 52% for genus level and 34% for species level [17]. During the past decade, biochemical methods are being replaced by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS), which is discussed in the following section.

Mycobacteria should be cultured in a biosafety level 3 or BSL2 + (BSL2 plus BSL 3 practices) laboratory due to the potential aerosolization of *Mycobacterium tuberculosis* (MTB). Mycobacteria were traditionally identified by biochemical methods. However, this process required weeks for the slowly growing mycobacterial species, including MTB. Improvements in time to identification were made with the implementation of high-performance liquid chromatography (HPLC), DNA probes, MALDI-TOF MS, and polymerase chain reaction (PCR) for MTB complex. Biochemical methods and HPLC are no longer recommended due to their poor performance and slow turnaround times compared with newer methods [22]. Although DNA probes provide a same day turnaround time once growth is detected, they are labor-intensive and only available for four species: MTB complex, *Mycobacterium avium* complex, *Mycobacterium kansasii*, and *Mycobacterium gordonae*. In addition to routine bacteria, MALDI-TOF MS is also replacing traditional methods for identification of mycobacteria.

Yeasts have historically been identified using biochemical methods, which require up to 72 hours. The phenotypic methods commonly employed by clinical laboratories for yeast identification have variable accuracy, with most methods performing well for common species; however, accuracy may decrease with infrequently encountered isolates [23–27]. One of the notable drawbacks of biochemical methods for yeast testing is the misidentification of *Candida auris* [28]. *C. auris* has gained international attention due to isolates frequently demonstrating multidrug resistance. Molds, on the other hand, have historically been identified using macroscopic and microscopic morphologies. The surface color and texture of the mold colony, the color of the reverse of the colony, and the microscopic morphology using lactophenol cotton blue (LPCB) tape preparation are combined to identify the mold. LPCB tape preparation is performed by touching a piece of clear tape to the top of the mold colony and applying it to a microscope slide with a drop of LPCB dye, which stains the fungal elements and allows visualization. MALDI-TOF MS is replacing biochemical methods for yeast. MALDI-TOF MS has also been applied to mold identification. However, there are some drawbacks specific to molds, which have resulted in limited usage compared with other organism groups and are discussed in the following section.

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry

MALDI-TOF MS is a rapid and inexpensive method for identification of bacteria, acid-fast bacilli (*Mycobacterium* spp.), yeasts, and molds in the clinical microbiology laboratory. MALDI-TOF MS uses proteomic profiling to assign and identification, which primarily includes ribosomal proteins due to their relative abundance. Thus MALDI-TOF MS is the closest identification to the gold standard of 16S ribosomal DNA sequencing. MALDI-TOF MS measures proteins in the range of 2–20 kilodaltons. Bruker's Biotyper system (Fig. 55.5A) and bioMerieux's VITEK MS (Fig. 55.5B) are the two commercially available MALDI-TOF MS instruments available in the United States for microorganism identification. The Bruker is a bench top model, while the VITEK MS is a larger floor model. Both systems have Food and Drug Administration (FDA)-cleared in vitro diagnostic (IVD) organism libraries as well as more comprehensive research use only (RUO) libraries. Although the majority of clinically encountered bacteria and yeast can be identified using the IVD libraries, many laboratories have validated RUO libraries for clinical use to provide complete coverage of organism identification. Mycobacteria and

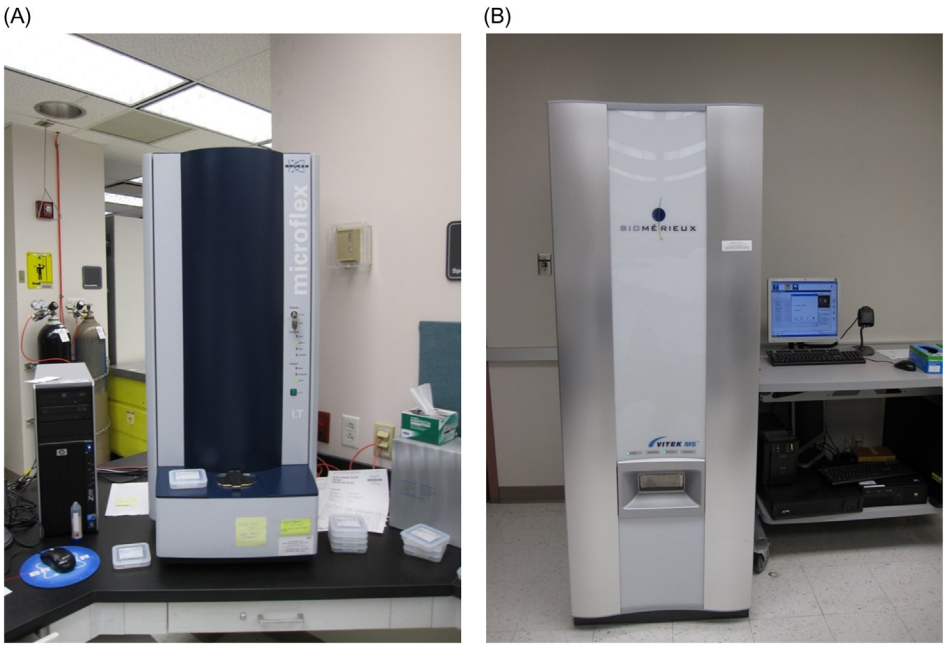


FIGURE 55.5 (A) Bruker's Biotyper system and (B) bioMerieux's VITEK MS are the two commercially available matrix-assisted laser desorption/ionization time-of-flight mass spectrometry instruments available in the United States for microorganism identification.

molds are included in the most recent VITEK MS IVD library v.3.0, while the Bruker has RUO mycobacterial and mold libraries that must be purchased separately from the bacterial library.

For bacteria and yeasts, a thin layer of organism is applied to the target plate using a toothpick or inoculation loop and 1 μL of formic acid overlay may be added, depending on the organism group. The organism ± formic acid is subsequently overlaid with matrix (α-cyano-4-hydroxycinnamic acid), allowed to dry, and subjected to MALDI-TOF analysis. The process is more complicated for mycobacteria and molds, which require extraction and inactivation prior to MALDI-TOF MS analysis. Of note, both commercially available laboratory automation systems are developing modules that will automate inoculation of MALDI-TOF targets. Once the organism's mass spectrum is obtained, it is compared with the manufacturer's library/database, and a confidence score is provided. The Bruker uses a logarithmic scale, where, per the manufacturer, a score ≥ 1.7 is sufficient for genus-level identification and a score of ≥ 2.0 is sufficient for species-level identification. Some laboratories have validated lower thresholds off-label for species-level identification. The VITEK MS system reports, on the other hand, identifications with a percent confidence score up to 99.9%.

MALDI-TOF MS has a turnaround time of approximately 40 minutes for identification of bacteria and yeasts [8,29]. This is a significant improvement compared with 1.5−72 hours for traditional biochemical methods [8,29]. In addition to a faster turnaround time, MALDI requires less organism mass. A single isolated colony is sufficient for MALDI-TOF MS analysis, while biochemical methods require sufficient colonies to create a 0.5−3 McFarland turbidity standard, depending on the organism group.

MALDI-TOF MS has demonstrated excellent performance for aerobic bacteria with genus- and species-level identification rates of 98.6% and 96.5% for Gram-positives and 98.5% and 96.8% for Gram-negatives [30]. One of the biggest improvements over biochemical methods is the identification of anaerobes. Genus- and species-level identification rates for anaerobes by MALDI-TOF MS have been reported to be 98.5% and 85.8%−97.4%, as compared with 71%−87% and 50%−60%, respectively, by biochemical methods [18−21,30]. Approximately 60% of College of American Pathologists (CAP)-accredited laboratories are using MALDI-TOF MS for the identification of bacteria (CAP DEX-B 2019 survey).

MALDI-TOF MS has a few limitations with regard to bacterial identification. It is unable to separate *E. coli* and *Shigella* spp., which are essentially the same organism. The technology also has challenges with discrimination and accurate detection of *Streptococcus pneumoniae* and *Streptococcus mitis/oralis* group due to the high degree of relatedness between the two organisms; however, this has improved with the release of updated libraries. MALDI-TOF MS typically does not identify or misidentifies select agents, such as *Brucella* spp., *B. anthracis, Burkholderia pseudomallei*, etc. even when they are present in the library being used [31]. Therefore, it is imperative that microbiology technologists are familiar with recognizing select agents and recognizing common misidentifications by MALDI-TOF MS. It is essential that technologists are

aware of the limitations and maintain competency for recognizing patterns and questioning things that do not fit. For microorganisms not identified by MALDI-TOF, DNA sequencing should be used if identification is necessary.

A recent CAP proficiency testing survey (E-B 2019) revealed that 20% of laboratories are using MALDI-TOF MS for identification of mycobacteria. MALDI-TOF MS performs well for mycobacteria, although more hand-on time is required than bacteria and yeast. The Bruker extraction/inactivation method requires approximately 1 hours, while the VITEK MS extraction method requires 30−60 minutes. However, a recent study published a method that requires only 10 minutes, decreasing the hands-on time and turnaround time [32]. Species-level identification rates are 84.7%−100% [32−34].

Notable drawbacks of MALDI-TOF are that it cannot differentiate members of the MTB complex nor can it differentiate *Mycobacterium abscessus* complex to subspecies-level (*M. abscessus* subsp. *abscessus, M. abscessus* subsp. *bolletii* and *M. abscessus* subsp. *massiliense*). The former is significant because *Mycobacterium bovis*, a member of the MTB complex, is intrinsically resistant to pyrazinamide, a first-line drug for tuberculosis treatment. The latter is significant, because *M. abscessus* subsp. *abscessus* and *M. abscessus* subsp. *bolletii* have a functional *erm* gene, encoding macrolide resistance, whereas *M. abscessus* subsp. *massilience* has a nonfunctional *erm* gene. MALDI-TOF MS also cannot differentiate *Mycobacterium chimaera* from *Mycobacterium intracellulare*. While this is normally not a concern, sequencing should be performed if the infection is associated with cardiothoracic surgery, as there is currently a global outbreak of *M. chimaera* associated with heater-cool units [35].

Currently, 48% of CAP-accredited laboratories are using MALDI-TOF MS for identification of yeasts (CAP F-B 2019 survey). Several studies have demonstrated species-level identification rates of 83%−99% (reviewed in [36]), and the majority of yeasts require formic acid overlay or extraction with formic acid and acetonitrile for successful identification. MALDI-TOF MS identification of molds, on the other hand, has been less widespread. While 48% of laboratories are using MALDI-TOF MS for yeasts, only 5% are using it for molds (CAP F-B 2019 survey). There are no molds in the most recent FDA-cleared library (IVD Claim 4) for Bruker, and there are 79 species in the recent VITEK MS library (v3.0). The RUO libraries for both instruments, however, contain numerous species. Although libraries are available, there are two issues with identification of molds. Mold colonies are not homogeneous the way bacterial, mycobacterial, or yeast colonies are. Protein expression varies across mold colonies, with a subset of proteins being more highly expressed at the inner portion of the colony, whereas another subset of proteins may be more highly expressed at the outer edge of the colony, etc. Thus the sampling location has a significant impact on which proteins are present in the spectra. Moreover, in the case of the Bruker, the original mold library was created using molds that had been grown in broth rather than on plates. Although the broth method standardizes protein expression, molds are not grown in broth in clinical laboratories.

The difficulty in mold identification is demonstrated by publications reporting 78%−83.5% identification to genus level and 54%−79% identification to species level (reviewed in [36]). Some laboratories have created their own user-developed libraries, either for all molds or for a particular subset of molds. User-developed libraries have demonstrated species-level identification rates of 89%−98% [36]; however, many labs do not have the capacity to create their own library. A recent study evaluating the VITEK MS v3.0 library demonstrated 86.6% correct identification to the species level [37]. If only the species present in the library were included, 91% were correctly identified to species level. These are the most promising data using a commercially developed library thus far. MALDI-TOF MS identification of molds is an ongoing area of research.

Beyond colony identification

To date, the majority of studies evaluating MALDI-TOF MS have focused on identification of organisms growing on solid media; a smaller number of studies have evaluated MALDI-TOF MS analysis on positive blood culture broth. An initial extraction is required when testing positive blood cultures to eliminate interfering substances, including blood and broth. Time to identification ranges from 30 minutes to 2 hours, depending on the method used. Of note, this method applies to bacteria and yeasts, which grow well in routine blood culture bottles. Mycobacteria and molds are rarely recovered in routine blood culture bottles and require alternative methods for detection. Correct identification rates vary based on which extraction method is utilized but appear most optimized with the use of a commercially available Sepsityper kit (Bruker-Daltonics) or saponin [38−42]. The majority of studies have evaluated the Bruker Biotyper for this analysis. One drawback is that MALDI-TOF cannot accurately identify multiple species in polymicrobial blood cultures, though it frequently identifies the predominant organism from mixed cultures [43−46]. The issue with polymicrobial specimens has also been reported with molecular methods [47,48].

MALDI-TOF MS has been shown to correctly identify 84%−99% of Gram-negative bacteria from positive blood culture broth, with most studies reporting rates from 90% to 99% (reviewed in [49]). The rate varies from

65% to 96% for Gram-positive bacteria, with most studies reporting identification rates of 80% [49]. In regard to the identification of yeasts, one study showed concordance rates of 95.9% for *Candida albicans* and 86.5% for non-*albicans Candida* species [50].

Rather than using lysis or extraction methods, some labs have opted to test microcolonies or "scum" growth. Positive blood culture broth is inoculated to solid media and incubated, and MALDI-TOF MS analysis is performed after 2−6 hours of incubation. Although the turnaround time is longer than the extraction methods, this method is less labor-intensive and fits better into a laboratory workflow. Species-level identification rates have been reported to be in the range of 81.8%−95.5% [49]. Anaerobes and slow growing bacteria have lower identification rates, and polymicrobial cultures cannot be differentiated after 2−6 hours of growth.

Direct identification from urine via MALDI has also been investigated. One study used diafiltration and concentration to obtain bacterial identification in 2−3 hours. Detection limits were 5×10^4 to 10^6 colony forming units/mL [51]. A second study inoculated urine into broth media, incubated the broth for 3 hours, and checked the density. If the density was ≥ 0.3 on the McFarland turbidity scale, the broth was centrifuged, and the bacterial pellet was used to spot the MALDI target. The study demonstrated 96.5% sensitivity and 71.4% specificity [52]. Neither workflow is ideal for a clinical microbiology laboratory, so MALDI-TOF MS of urine cultures is not widely performed. In addition, urine cultures have a high rate of contamination and are often polymicrobial in nature, further limiting the utility of this method.

Point-of-care microbiology

Point-of-care (POC) testing generally refers to testing that is performed near the site of patient care (such as a physician office, clinic, or hospital unit) by nonlaboratory trained individuals. POC tests often have a rapid turnaround time, providing actionable results that facilitate immediate patient management decisions. This advantage is particularly pronounced in the emergency department and in cases where the traditional laboratory is off-site. A more expanded definition of POC testing may include testing performed in small laboratories that serve as a satellite laboratory for a main central laboratory. Centralization of laboratory testing, where a large centralized laboratory performs the majority of testing for hospitals and clinics within a region, has led to increased turnaround times for results when transport times are taken into account. Smaller satellite laboratories may be staffed by general laboratory technologists without specialized microbiology training. Testing performed in these types of laboratories may confer the turnaround time benefits of POC testing by eliminating or reducing transit times while offering an expanded menu of assays. In this context, many rapid sample-to-answer microbiology tests may be considered POC tests, such as syndromic multiplex molecular testing platforms. For remote clinics or hospitals, simple and rapid POC assays are essential for optimal patient management.

Most POC tests receive waived status by the United States FDA. Laboratories or clinics exclusively performing waived tests are required to obtain a Certificate of Waiver, which is issued by the Centers for Medicare and Medicaid Services under Clinical Laboratory and Improvement Amendments of 1988 (CLIA). Waived tests are considered low risk for errors, are performed using simple and straightforward protocols, require minimal instrument maintenance, and require minimal training for test interpretation. Although these assays are simple, appropriate quality control and user training is essential to preventing diagnostic errors.

POC microbiology testing has historically been limited to lateral flow immunoassays (LFAs). LFA tests are performed on a lateral flow strip and are often double-antibody sandwich assays. A capture antibody is immobilized on the membrane and the clinical specimen is flowed over the strip. The target antigen, if present, is bound by conjugate-labeled antibodies, such as colloidal gold-labeled antibodies, and the target antigen−antibody complex binds the immobilized capture antibody, producing a visible line. The sample flows further along the membrane where excess conjugate binds the control antibodies, producing a visible control line. For some assays, digital analyzers have been introduced to eliminate the subjective nature of visual interpretation of LFA assays. Examples of LFA POC tests in use include assays for the diagnosis of *S. pyogenes* [Group A *Streptococcus* (GAS)], infectious mononucleosis, influenza A/B viruses, respiratory syncytial virus (RSV), HIV-1 and HIV-2, and adenovirus.

Although LFA assays are rapid, inexpensive, and simple, these assays have variable clinical performance and often suffer from suboptimal sensitivity. Metaanalysis of the diagnostic accuracy of rapid antigen tests for GAS found a sensitivity of 86% and specificity of 95%, although the reported sensitivities in the included studies were highly variable, ranging from 44% to 98% [53]. Owing to the low sensitivity of these tests, negative rapid GAS tests are reflexed to culture for confirmation testing.

Rapid influenza diagnostic tests (RIDTs) are immunoassays that detect influenza viral antigens. Several metaanalyses have evaluated the sensitivity of RIDTs, reporting a pooled sensitivity of 51%−68% and lower sensitivity for adult patients compared with pediatric patients due to the relatively lower viral load [54−56]. When seasonal influenza activity is high, the negative predictive value of RIDTs is low, due to the substantial

number of falsely negative results. Because of these poor performance characteristics, the Centers for Disease Control and Prevention recommends that all negative samples be tested by a more sensitive assay, although adherence to this recommendation is likely low. Further, antiviral therapy is commonly not withheld based on a negative RIDT alone, particularly if there is a high clinical suspicion for influenza illness. Conversely, when influenza activity is low to moderate, the positive predictive value of RIDTs is poor. Due to low sensitivity for the detection of influenza viruses, especially novel or emerging influenza viruses, in 2017 the FDA reclassified RIDTs from class I to class II devices, which requires these devices meet minimum criteria for sensitivity and specificity [57].

Development of POC molecular techniques has the potential to meet the clinical need for highly sensitive, rapid POC microbiology diagnostics. The first POC molecular test to be granted a CLIA waiver was the Alere i influenza A & B test (Alere Inc.), which received waived status in 2015. Subsequently, Abbott purchased Alere and renamed the Alere i to the infectious disease (ID) NOW platform. It is CLIA-waived for influenza A/B virus, RSV, and GAS testing, with results reported in 2−15 minutes. The ID NOW system uses strand displacement amplification technology, where target nucleic acid is amplified in an isothermal reaction using a nicking endonuclease and a strand-displacing DNA polymerase in addition to target primers. Fluorescently labeled molecular beacon probes are used to monitor amplification in real time. Three additional molecular platforms have subsequently received CLIA-waived status. The Cobas Liat (Roche Diagnostics) is CLIA-waived for influenza A/B virus, RSV, and GAS testing, with results in 15−20 minutes. Amplification occurs using real-time PCR technology and results are interpreted based on real-time detection of fluorescently labeled probes. The GeneXpert Xpress platform (Cepheid) is CLIA-waived for influenza/RSV and GAS testing, with results in 18−30 minutes. The Xpress system uses real-time PCR and fluorescently labeled probes for the amplification and detection of target nucleic acid. The Accula system (Mesa Biotech) is CLIA-waived for influenza and RSV detection, with results in 30 minutes. The Accula system uses reverse-transcription PCR technology to amplify target RNA, which is then complexed to oligonucleotide probes conjugated to dye-labeled microspheres. Using technology similar LFAs, the amplicon-microsphere complex is flowed across a detection membrane where complementary oligonucleotide capture probes are bound. Hybridization of capture probes to the amplicon-microsphere complex is visualized on the membrane as a colored line. Additional molecular POC assays are currently under development and in clinical trials.

Molecular POC tests have demonstrated clinical performance characteristics similar to molecular testing performed in the clinical microbiology laboratory. GAS molecular POC assays have a reported sensitivity of 99%−100% and specificity of 91%−97% when compared with culture [58,59]. POC molecular assays for RSV and influenza A/B viruses also demonstrate high sensitivity. Reported sensitivity of the Liat is >97% for influenza A virus, influenza B virus, and RSV targets [60−63]. Similarly, evaluation of the GeneXpert influenza A/B assay demonstrated high sensitivity (>96%) and specificity (>97%) for both influenza A and influenza B viruses [63,64]. Studies of the Alere i have reported lower sensitivity for all targets when compared with the Liat, with overall influenza A/B sensitivity as low as 64% in one study, although the majority of studies report sensitivities >90% [60,63−66]. Metaanalysis of the reported diagnostic accuracy of influenza POC nucleic acid amplification tests found a pooled sensitivity 92% for influenza A virus and 95% for influenza B virus [67]. This was a marked increase compared with the pooled sensitivity of newer generation influenza antigen immunoassays, which had a pooled sensitivity of 80% for influenza A virus and 77% for influenza B virus.

Risk of contamination is a significant concern for molecular POC assays, as amplicon contamination can lead to false-positives and pseudooutbreaks. Because these tests are performed by nonlaboratory staff that are inexperienced in performing highly sensitivity laboratory assays, proper training and quality control, such as frequent testing of the setup area for target contamination, are essential. In addition, improved test utilization education and stewardship strategies are a focus across all areas of laboratory medicine. As increasingly expensive and sensitive tests are introduced, education on which patients are appropriate to test will be essential. For example, overdiagnosis of pediatric patients colonized with GAS is a concern, as up to 26% of healthy children may be colonized with GAS [68,69]. Testing only patients with clinically significant signs and symptoms of GAS is imperative to prevention of overdiagnosis.

A drawback of molecular POC testing is increased cost per test compared with traditional antigen tests. Benefits that may offset cost increases include the potential for improved antibiotic and antiviral use. In the cases of GAS, patients that tested antigen-negative but culture-positive would potentially receive results and antibiotic therapy 1−2 days sooner if molecular testing was used. In addition, molecular GAS POS testing does not require reflex of negative results to culture, reducing the workflow burden in the clinical microbiology laboratory and simplifying the overall operational process. In the cases of influenza A/B testing, the increased negative predictive value of molecular testing over antigen testing during influenza virus season may facilitate appropriate use of antiviral medications and judicious use of antibiotics.

Molecular POC platforms will allow for expansion of the POC microbiology test menu. Numerous assays are currently in development and many are currently CE-marked for use in European laboratories. Promising new assays include HIV-1 viral load testing and diagnostic assays for sexually transmitted infections. In 2019 the first assay for molecular POC STI detection received FDA 510(k) clearance. The binx *io* platform is a rapid molecular assay for the diagnosis of *Chlamydia trachomatis* and/or *Neisseria gonorrhoeae* from clinician or self-collected vaginal swab specimens, with results in less than 30 minutes.

Introduction of molecular microbiology into the POC testing environment represents a paradigm shift and has led to increased versatility, scope, and availability of POC rapid testing. POC molecular testing is also accompanied by an increase in cost per test, which may limit uptaking these assays, especially if viewed independent of changes to cost utilization of other resources, such as administration of antibiotics.

Syndromic-based multiplex molecular testing

Introduction

Multiplex testing was developed for the detection of pathogens associated with clinical syndromes of the bloodstream, respiratory, gastrointestinal, and central nervous systems (CNS). Since their introduction, these panels have become increasingly common with improved workflows and decreased turnaround times. Syndromic-based testing has streamlined ordering workflows, as providers need to select only one test order to analyze the most common organisms causing a specific syndrome. Moreover, some platforms offer flexibility in ordering, allowing the provider to order more specific targets based on clinical presentation, which can help limit unnecessary testing and costs. Molecular syndromic testing is more rapid than conventional microbiology methods, reducing time to diagnosis and, if the result is acted upon, potentially reducing time to appropriate therapy.

Early multiplex testing platforms, such as the Luminex xTAG, had cumbersome workflows and testing was often batched and performed once per day or once per shift. Subsequently, sample-to-answer platforms were developed and have replaced many of the older platforms. Sample-to-answer platforms require minimal set time, leading to simplified workflow as compared with traditional microbiology methods. A major limitation to adoption of multiplex testing is cost, as these panels incur an increased cost to both the laboratory as well as the patient compared with conventional methods. Institutional savings outside the clinical laboratory may be observed if utilization of these panels results in reduced antibiotic use, decreased length of stay, and promotion of appropriate infection prevention and control isolation precautions. In order to maximize institutional and patient benefits, multiplex testing should optimally be performed 24 hours per day and 7 days per week. In addition, test results are most optimally acted upon when accompanied by an antimicrobial stewardship intervention, either via direct contact with a member of an institutional antimicrobial stewardship team or via laboratory comments appended to results. Finally, test utilization strategies specific to each panel and patient population should be adopted to minimize inappropriate costly testing and target these assays to patients who will see the most benefit. Antimicrobial stewardship and test utilization opportunities specific to each syndromic panel are discussed below.

Currently, sample-to-answer multiplex testing platforms are commercially available via several vendors, including BioFire Diagnostics, the Luminex Corporation, and GenMark Diagnostics. GenMark panels are offered on the ePlex system, with a throughput of up to 96 samples within 8 hours. Following an amplification step, target DNA is hybridized to a signal probe, which reacts with a capture probe attached to a gold electrode. Application of voltage leads to electrochemical detection of the target DNA/signal probe complex. Current panels available on the ePlex instrument include respiratory pathogen and blood culture identification (BCID) panels. BioFire panels are offered on the FilmArray system. The FilmArray 2.0 is a single-bay instrument that is scalable by connecting up to eight instruments to a single processing computer. The FilmArray Torch is a fully integrated system with up to 12 modules in a tower configuration. In the FilmArray cartridge, nucleic acid is extracted and purified, followed by a nested PCR reaction and then a single-plex reaction for each target. Target detection is determined by endpoint melting curve analysis. BioFire currently has the largest menus of available panels, with testing available for infections of the upper respiratory, lower respiratory, gastrointestinal, and CNSs, as well as BCID. Luminex panels are performed on the Verigene system, consisting of a Verigene Processor and a Verigene reader. Specimen processing and target detection is performed on the Verigene Processor using a microarray format and gold nanoparticle probe-based technology. The test cartridge is then read by the Verigene reader, which offers scalable reading. Syndromic panels available on the Verigene system include respiratory, gastrointestinal, and BCID panels.

Respiratory panels

The differential diagnosis for upper respiratory infections is often broad, with nonspecific clinical signs and symptoms. Most illness is cause by viral infections, which do

not have targeted individual FDA-approved/cleared assays. Multiplex molecular testing for respiratory pathogens was first approved in 2008. Most panels target common respiratory viruses, with some including atypical bacterial targets (e.g., *Chlamydophila pneumoniae, Mycoplasma pneumoniae*, and *Bordetella* species). NP swabs are the only approved specimen type for these panels, although many laboratories have performed validation studies to allow for testing of lower respiratory specimens such as bronchoalveolar fluid. Several multiplex panels designed to detect respiratory pathogens have been FDA-cleared, including more traditional assays such as Luminex xTAG Respiratory Viral Panel (RVP; Luminex Corporation), the Luminex xTAG RVP Fast (Luminex Corporation), and the Luminex xTAG Respiratory Pathogen Panel (RPP; Luminex Corporation); newer sample-to-answer assays such as FilmArray Respiratory Panel (RP; BioFire Diagnostics) and RP2; Verigene Respiratory Pathogens Flex test (Luminex Corporation); and ePlex RPP (GenMark Diagnostics). Panels include 8−20 respiratory pathogen targets and turnaround time ranges from 1 to 8 hours.

Clinical evaluations of syndromic respiratory panels have demonstrated overall sensitivities >85% and specificities >99% [70−73]. However, suboptimal performance characteristics have been described for specific targets on individual panels. In one study, the FilmArray RP exhibited a sensitivity of 57% for adenovirus and 77% for influenza B virus when compared with a reference standard [70]. A redesign of the FilmArray RP, the RP2, has improved sensitivity for both adenovirus and influenza B virus [73]. The xTAG RVP Fast demonstrated <50% sensitivity for the detection of influenza B virus [70,74]. The xTAG-RPP demonstrated lower sensitivity for human coronaviruses OC43 (67%) and HKU1 (67%) than other assays [72]. Low sensitivity for viral pathogens with treatment options, such as influenza B virus, may negatively affect patient outcome; for this reason, a higher sensitivity assay should be used for the primary diagnosis of these viruses. For viruses not historically detected by traditional testing and for which no targeted therapy is available (such as coronaviruses), the impact of missed diagnosis is unclear. Nonetheless, many of these older assays have been supplanted by newer versions of sample-to-answer assays with improved overall sensitivity.

Implementation of multiplex respiratory testing has demonstrated clinical and financial benefits. Studies have documented a decreased mean time to diagnosis compared with conventional testing [75−77]. In addition, because multiplex panels include pathogens for which targeted testing is not routinely available (e.g., coronaviruses), use of multiplex panels increased the overall number of patients with a microbiological diagnosis [76]. Additional benefits include decreased use of antibiotics in viral illness, decreased admission rates, and decreased length of stay [75,76]. Although clinical benefits are promising, cost of implementation of multiplex panels is a significant concern. Mahony et al. [78] performed a cost evaluation of the xTAG RVP assay compared with conventional methods and determined that multiplex testing was cost-effective if the prevalence of respiratory viral illness was >11%. Availability and turnaround time of targeted assays in addition to the specific patient population (e.g., immunocompromised, pediatric, etc.) must be considered when evaluating the utility of multiplex respiratory pathogen tests.

The FilmArray Pneumonia (PN) panel received FDA approval/clearance in 2018. The PN panel includes targets for eight respiratory viruses, 18 bacteria, and seven antimicrobial resistance markers. This panel is approved for bronchoalveolar lavage (BAL), mini-BAL, sputum, and endotracheal aspirates. For 15 of the 18 bacterial targets, the target is reported as a semiquantitative result, which allows for interpretation of the results in a similar manner as traditional quantitative culture. Positive targets are reported within a 1 log bin, ranging from $<10^4$ to $\geq 10^7$ copies/mL. At the time of this writing, no peer-reviewed studies have been published on the performance of the PN panel.

Gastroenteritis panels

Traditional diagnostic testing for infectious diarrhea has included bacterial culture, antigen detection, microscopy, and targeted molecular tests. These assays require that clinicians select the appropriate test and misordering may lead to missed diagnoses. In addition, traditional testing for gastroenteritis (GI) pathogens in the microbiology laboratory is time-consuming, often requiring multiple days, and involves complicated workflows. Detection of some pathogens, such as *Giardia lamblia*, may require collection of multiple samples for maximum sensitivity. Advantages of syndromic panel testing for the diagnosis of diarrheal illness include decreased turnaround time, increased sensitivity, and, in some cases, decreased cost. Three multiplex assays are currently FDA-approved/cleared for the detection of enteric pathogens: Luminex xTAG Gastrointestinal Pathogen Panel (GPP; Luminex Corporation), FilmArray Gastrointestinal Panel (GIP: BioFire Diagnostics), and Verigene Enteric Pathogens Panel (Luminex Corporation). These multiplex tests detect 9−22 targets and on-instrument time is 1−5 hours. All panels detect the most common bacterial pathogens associated with GI (*Campylobacter* spp., *Salmonella* spp., *Shigella* spp., *Vibrio* spp., *Yersinia enterocolitica*, and Shiga toxin). The Luminex GPP and FilmArray GIP additionally detect enterotoxigenic *E. coli, E. coli* O:157, and *C. difficile*. The FilmArray GIP is the only assay that

detects enteropathogenic and enteroaggregative *E. coli* as well as *Plesiomonas shigelloides*. No multiplex GI panel detects *Aeromonas* species, an additional possible bacterial cause of infectious diarrhea. Of the viral causes of infectious diarrhea, all panels detect norovirus GI/II and rotavirus A. The Luminex GPP also detects adenovirus 40/41, while the FilmArray GIP also detects adenovirus 40/41, astrovirus, and sapovirus. Finally, the Luminex GPP and FilmArray GIP also include parasitic species, with the GPP detecting *Cryptosporidium* species, *Entamoeba histolytica*, and *G. lamblia* and the GIP detecting *Cryptosporidium* species, *E, histolytica*, *G. lamblia*, and *Cyclospora cayetanensis*. Stool submitted in Cary-Blair transport media is the acceptable specimen for all panels.

Multiplex panels offer significant sensitivity increases over conventional testing methods. Across multiple studies, all panels detect an enteric pathogen in 20%−50% of specimens, compared with detection of a pathogen in 8%−18% of samples tested by traditional methods [79−82]. Increased overall detection is likely due to increased sensitivity for specific targets as well as increased number of targets tested. The most commonly detected organisms are *C. difficile*, enteropathogenic *E. coli*, *Salmonella* species, norovirus, rotavirus, sapovirus, and *Cryptosporidium* species [79−82]. Sensitivity of all targets is generally >90%, although reduced sensitivities have been reported for *Salmonella* and *Y. enterocolitica* for the Luminex GPP and *Campylobacter*, *Salmonella*, and rotavirus for the Verigene EP [81,83,84]. Evaluation of multiplex molecular GI panels has repeatedly demonstrated that detection of >1 target is more frequent than previously appreciated. Codetections were observed in 16% of specimens tested by the BioFire GI Panel [79]. Among specimens with >1 targets detected, enteroaggregative and enteropathogenic *E. coli* were commonly detected [79−81].

Implementation of multiplex testing for the detection of GI pathogens has been associated with financial, clinical, and infection control benefits. In addition to increases in sensitivity and detection of coinfections, results are available in <6 hours when GI panel testing is performed, in contract with culture-based methods which require several days before a result is available. Although multiplex testing is associated with increased costs to the clinical laboratory when compared with traditional testing methods, in hospitalized patients, these costs may be offset by reduced overall institutional costs [85]. Improvement in utilization of patient isolation precautions associated with implementation of GI panel testing, including both increasing numbers of patients in appropriate isolation as well as discontinuing isolation precautions for patients with negative GI panel results, has important cost, patient satisfaction, and infection control implications [85,86]. Consideration should be given to test utilization and interpretation strategies when implementing syndromic GI testing for optimal cost and clinical utility.

Blood culture identification panels

Early antimicrobial administration is an important predictor of morbidity and mortality associated with sepsis. Diagnosis of bacteremia and fungemia is predicated on appropriate collection of blood cultures, with at least two sets of aerobic and anaerobic blood culture bottles filled with 10 mL of blood per bottle or 5 mL of blood per bottle for pediatric bottles. Blood cultures are incubated for five days on a continuously monitored blood culture instrument. When a bottle signals positive, a Gram stain is performed and the results are immediately called to the clinical care team. The blood culture broth is then inoculated to solid media and organism identification and antimicrobial susceptibility testing is performed on culture growth, requiring an additional 18−48 hours. From the time of initial positivity, workup of positive blood culture bottles requires 48−72 hours before the report is finalized. Multiplex molecular testing of positive blood culture bottles may decrease both the time to organism identification and time to preliminary antimicrobial susceptibility results.

Currently, several assays have received FDA approval/clearance for testing positive blood culture bottles. The FilmArray BCID panel tests for 19 bacterial targets (Gram-positive and Gram-negative), five yeast targets, and four resistance genes. The Luminex Verigene assay is divided into two panels based on Gram stain results. The Gram-positive blood culture (BC-GP) panel tests for 12 Gram-positive bacterial targets and three resistance genes. The Gram-negative blood culture (BC-GN) panel contains eight Gram-negative bacterial targets and six resistance markers. The GenMark ePlex system offers three panels. The Gram-positive panel contains 20 bacterial targets and four resistances gene; the Gram-negative panel tests for 21 bacterial targets and six resistance genes; and the fungal panel tests for 15 fungal targets. Finally, the Accelerate Pheno system (Accelerate Diagnostics) uses gel electrofiltration and fluorescence in situ hybridization for the identification of six Gram-positive bacteria, eight Gram-negative bacteria, or two yeast targets within 90 minutes and subsequent phenotypic susceptibility testing within 7 hours. The Accelerate Pheno system is the only multiplex system that offers phenotypic susceptibility testing.

Performance characteristics of multiplex molecular testing have been evaluated across multiple studies. Implementation of molecular testing of positive blood culture bottles reduced time to organism identification by 20−30 hours [87,88]. In monomicrobial positive blood

culture bottles, an organism was correctly identified in 90%–100% of bottles [48,87–89]. In addition, select resistance genes have been incorporated into these panels for a preliminary indication of antimicrobial resistance, including *mecA* detection for the preliminary identification of MRSA species, *vanA/B* detection for the preliminary identification of VRE species, $bla_{CTX\text{-}M}$, the most commonly detected extended spectrum beta-lactamase in *Enterobacterales*, and the five most common carbapenemase genes detected in the United States. Assays from each manufacturer have different combinations of genetic resistance markers. Assessment of performance characteristics of the susceptibility markers on each panel have shown high rates of agreement between assay detection and conventional testing and/or sequencing [47,48,89].

False-positive results have been reported using the FilmArray BCID, possibly due to microbial DNA contamination of blood culture broth. *P. aeruginosa* DNA has been detected in bioMerieux BacT/Alert standard anaerobic bottles when tested by the FilmArray BCID [87]. Moreover, false identification of *Candida parapsilosis* and *Proteus* species has been reported using BD Bactec blood culture bottles when testing by the FilmArray BCID. The Verigene BC-GN and BC-GP assays, on the other hand, have not been affected by contaminating DNA in blood culture bottles, because the assays do not include an amplification step. Due to the possibility of false-positive target detection, results from molecular testing should be correlated with the bottle Gram stain and discordant results warrant further testing and investigation.

Numerous studies have evaluated the clinical impact of implementation of rapid BCID. Differences in study population, individualized institutional practices, and study design lead to differences in outcome metrics across studies. Decreased time to appropriate antibiotic therapy was observed in most studies following implementation of rapid blood culture panels [90–96]. In addition, some but not all studies have demonstrated decreased length of hospital stay and decreased overall institution costs [90,94,96]. No mortality benefit was noted following implementation in most clinical studies; however, a few studies reported a statistically significant reduction in 30-day mortality rate [95,96].

Maximum impact of these panels is appreciated when results are coupled with antibiotic stewardship interventions, providing real-time guidance to clinicians to promote optimization of antibiotic choice and dose [97,98]. Antibiotic stewardship interventions may include real-time communication of BCID results from the microbiology laboratory to a member to the antimicrobial stewardship team (such as an ID pharmacist or ID physician). In such interventions, real-time antibiotic decision support is then provided to the primary clinical team by an ID expert. While real-time in-person interventions are likely to result in optimal antibiotic therapy, institutions may not have sufficient resources to implement fully these labor-intensive interventions. Alternatively, laboratories may release BCID testing results with evidence-based antibiotic stewardship recommendations in the form of comments appended to test results. Such comments should be carefully crafted in consultation with ID specialists. Although BCID reporting comments provide decision support and are more easily implemented compared with real-time expert advice, comments do not provide individualized guidance and are easily overlooked by physicians. Ultimately, improved communication between primary clinical providers, antibiotic experts, and the clinical microbiology laboratory is essential for optimal utilization of BCID panel results and improved patient care.

Meningitis and encephalitis panel

Infections of the CNS are associated with significant morbidity and mortality. Establishing a specific diagnosis is integral to administering appropriate therapy, as the etiologic agents of meningitis and encephalitis vary. Clinical presentation may be nonspecific, with many patients experiencing headache, alternated mental status, and potentially nuchal rigidity. WBC count and differential in cerebrospinal fluid, in addition to protein and glucose concentrations, may be indicative of the infecting agent. In the cases of bacterial meningitis, WBC counts are typically elevated with a neutrophilic predominance. Moreover, CSF protein concentrations are usually elevated and glucose concentrations are usually below normal limits. In the cases of viral meningitis, WBC counts are also elevated, with a lymphocytic predominance, although neutrophils may be present early in infection. Protein concentration is normal to elevated and glucose concentrations are generally normal. In the cases of fungal or tubercular CNS infections, WBC counts are elevated with lymphocytosis, protein concentrations are normal to elevated, and glucose concentration is normal to low. Traditional microbiologic testing of CSF is necessary to determine the infecting agent. Bacterial meningitis is generally diagnosed by Gram stain and bacterial culture. Viral meningitis requires molecular testing of CSF for the diagnosis of enterovirus and herpes simplex virus 1 and 2 (HSV-1 and HSV-2). In some cases of viral CNS infections, such as in cases of vector-borne encephalitis, serological testing may be warranted. Tubercular and fungal CNS infections are diagnosed by molecular and culture-based methods and, in the case of cryptococcal meningitis, by antigen detection in the CSF. Molecular detection of CNS pathogens may allow for a more rapid diagnosis, especially in the cases of bacterial meningitis. In addition, molecular testing is less likely to be affected

by prior antimicrobial therapy that may affect culture sensitivity.

At this time, only one panel has received FDA approval/clearance for the diagnosis of CNS infections, the FilmArray Meningitis/Encephalitis (ME) panel (BioFire Diagnostics). This panel tests for 14 pathogens, including enterovirus, human parechovirus, HSV-1 and -2, cytomegalovirus, human herpesvirus 6, varicella zoster virus, *E. coli* K1, *Haemophilus influenzae, Listeria monocytogenes, N. meningitidis, Streptococcus agalactiae*, *S. pneumoniae*, and *Cryptococcus neoformans/gattii*. Several studies have evaluated the performance of the ME Panel and found high overall positive and negative agreement rates. In a large prospective evaluation of CSF specimens, the authors found 84.4% and >99.9% positive and negative agreement between the FilmArray ME Panel and conventional microbiology testing [99]. In a smaller study of pediatric patients, authors found 92.5% positive agreement and 96.2% negative agreement [100].

False-positive results from these panels are a matter of concern and prevention of contamination during specimen collection and handling is integral to appropriate utilization of the FilmArray ME Panel. Several studies have documented concerning rates of *S. pneumoniae* false-positive detection [99,101]. For these reasons, many laboratories have placed testing restrictions on these panels. In some cases, a laboratory director may opt to evaluate CSF parameters prior to performing the FilmArray ME Panel. Alternatively, some laboratory directors may evaluate all positive results prior to release to clinicians. False-negative results are also concerning. Cases of false-negative results for cryptococcal meningitis have been documented when compared with cryptococcal culture results [101,102]. Finally, some studies have reported false-negative HSV-1 and HSV-2 results, possibly due to viral burden below the panel limit of detection in Refs. [100] and [101].

Additional studies are necessary to determine the clinical and economic benefits of syndromic panel testing for the diagnosis of meningitis and encephalitis. Pre-/postintervention studies evaluating the impact of FilmArray ME Panel testing on length of stay and duration of antibiotic treatment have shown variable results, with reduced length of stay documented in one study but not observed in another [103,104]. Similar to outcome studies associated with other syndromic panels, it is likely that antibiotic stewardship intervention at the time of result reporting will be helpful in realizing the maximum impact of these results.

Conclusion

The clinical microbiology laboratory has undergone a period of rapid growth and evolution during the past 10–15 years. Significant strides in advancement continue to be made as more rapid and automated technologies obtain FDA-clearance.

References

[1] J.M. Miller, M.J. Binnicker, S. Campbell, K.C. Carroll, K.C. Chapin, P.H. Gilligan, et al., A guide to utilization of the microbiology laboratory for diagnosis of infectious diseases: 2018 update by the Infectious Diseases Society of America and the American Society for Microbiology, Clin. Infect. Dis. 67 (2018) e1–e94.

[2] J.M. Miller, S.A. Miller, A Guide to Specimen Management in Clinical Microbiology, third ed., ASM Press, 2017.

[3] C. Dempsey, E. Skoglund, K.L. Muldrew, K.W. Garey, Economic health care costs of blood culture contamination: a systematic review, Am. J. Infect. Control. 47 (2019) 963–967.

[4] J. Iversen, G. Stendal, C.M. Gerdes, C.H. Meyer, C.O. Andersen, N. Frimodt-Moller, Comparative evaluation of inoculation of urine samples with the Copan WASP and BD Kiestra InoqulA instruments, J. Clin. Microbiol. 54 (2016) 328–332.

[5] A. Croxatto, K. Dijkstra, G. Prod'hom, G. Greub, Comparison of inoculation with the InoqulA and WASP automated systems with manual inoculation, J. Clin. Microbiol. 53 (2015) 2298–2307.

[6] C. Quiblier, M. Jetter, M. Rominski, F. Mouttet, E.C. Bottger, P.M. Keller, et al., Performance of Copan WASP for routine urine microbiology, J. Clin. Microbiol. 54 (2016) 585–592.

[7] A. Cherkaoui, G. Renzi, N. Vuilleumier, J. Schrenzel, Copan WASPLab automation significantly reduces incubation times and allows earlier culture readings, Clin. Microbiol. Infect. 25 (2019) 1430.e5–1430.e12. Available from: https://doi.org/10.1016/j.cmi.2019.04.001.

[8] T. Theparee, S. Das, R.B. Thomson Jr., Total laboratory automation and matrix-assisted laser desorption ionization-time of flight mass spectrometry improve turnaround times in the clinical microbiology laboratory: a retrospective analysis, J. Clin. Microbiol. 56 (2018).

[9] P. Froment, H. Marchandin, P. Vande Perre, B. Lamy, Automated versus manual sample inoculations in routine clinical microbiology: a performance evaluation of the fully automated InoqulA instrument, J. Clin. Microbiol. 52 (2014) 796–802.

[10] M. Graham, L. Tilson, R. Streitberg, J. Hamblin, T.M. Korman, Improved standardization and potential for shortened time to results with BD Kiestra total laboratory automation of early urine cultures: a prospective comparison with manual processing, Diagn. Microbiol. Infect. Dis. 86 (2016) 1–4.

[11] N.T. Mutters, C.J. Hodiamont, M.D. de Jong, H.P. Overmeijer, M. van den Boogaard, C.E. Visser, Performance of Kiestra total laboratory automation combined with MS in clinical microbiology practice, Ann. Lab. Med. 34 (2014) 111–117.

[12] M.L. Yarbrough, W. Lainhart, A.R. McMullen, N.W. Anderson, C.D. Burnham, Impact of total laboratory automation on workflow and specimen processing time for culture of urine specimens, Eur. J. Clin. Microbiol. Infect. Dis. 37 (2018) 2405–2411.

[13] M.L. Faron, B.W. Buchan, C. Coon, T. Liebregts, A. van Bree, A.R. Jansz, et al., Automatic digital analysis of chromogenic media for vancomycin-resistant-enterococcus screens using Copan WASPLab, J. Clin. Microbiol. 54 (2016) 2464–2469.

[14] M.L. Faron, B.W. Buchan, C. Vismara, C. Lacchini, A. Bielli, G. Gesu, et al., Automated scoring of chromogenic media for detection of methicillin-resistant Staphylococcus aureus by use of WASPLab image analysis software, J. Clin. Microbiol. 54 (2016) 620–624.

[15] A. Croxatto, R. Marcelpoil, C. Orny, D. Morel, G. Prod'hom, G. Greub, Towards automated detection, semi-quantification and identification of microbial growth in clinical bacteriology: a proof of concept, Biomed. J. 40 (2017) 317–328.

[16] CLSI. Clinical Laboratory and Standards Institute (CLSI) Guideline MM18: Interpretive Criteria for Identification of Bacteria and Fungi byTargeted DNA Sequencing, second ed., 2018, Clinical and Laboratory Standards Institute, PA, USA.

[17] R.T. Saffert, S.A. Cunningham, S.M. Ihde, K.E. Jobe, J. Mandrekar, R. Patel, Comparison of Bruker biotyper matrix-assisted laser desorption ionization-time of flight mass spectrometer to BD Phoenix automated microbiology system for identification of gram-negative bacilli, J. Clin. Microbiol. 49 (2011) 887–892.

[18] M.J. Barba, A. Fernandez, M. Oviano, B. Fernandez, D. Velasco, G. Bou, Evaluation of MALDI-TOF mass spectrometry for identification of anaerobic bacteria, Anaerobe. 30 (2014) 126–128.

[19] L. Blairon, M.L. Maza, I. Wybo, D. Pierard, A. Dediste, O. Vandenberg, Vitek 2 ANC card versus BBL crystal anaerobe and RapID ANA II for identification of clinical anaerobic bacteria, Anaerobe. 16 (2010) 355–361.

[20] E.H. Lee, J.E. Degener, G.W. Welling, A.C. Veloo, Evaluation of the Vitek 2 ANC card for identification of clinical isolates of anaerobic bacteria, J. Clin. Microbiol. 49 (2011) 1745–1749.

[21] B. Rodriguez-Sanchez, L. Alcala, M. Marin, A. Ruiz, E. Alonso, E. Bouza, Evaluation of MALDI-TOF MS (Matrix-Assisted Laser Desorption-Ionization Time-of-Flight Mass Spectrometry) for routine identification of anaerobic bacteria, Anaerobe. 42 (2016) 101–107.

[22] CLSI. Clinical and Laboratory Standards Institute (CLSI) Guideline M48: Laboratory Detection and Identification of Mycobacteria, second ed., 2018, Clinical and Laboratory Standards Institute, PA, USA.

[23] D.J. Hata, L. Hall, A.W. Fothergill, D.H. Larone, N.L. Wengenack, Multicenter evaluation of the new VITEK 2 advanced colorimetric yeast identification card, J. Clin. Microbiol. 45 (2007) 1087–1092.

[24] J. Meletiadis, M. Arabatzis, M. Bompola, K. Tsiveriotis, S. Hini, E. Petinaki, et al., Comparative evaluation of three commercial identification systems using common and rare bloodstream yeast isolates, J. Clin. Microbiol. 49 (2011) 2722–2727.

[25] M. Castanheira, L.N. Woosley, D.J. Diekema, R.N. Jones, M.A. Pfaller, Candida guilliermondii and other species of candida misidentified as Candida famata: assessment by vitek 2, DNA sequencing analysis, and matrix-assisted laser desorption ionization-time of flight mass spectrometry in two global antifungal surveillance programs, J. Clin. Microbiol. 51 (2013) 117–124.

[26] S. Vijgen, S. Nys, R. Naesens, K. Magerman, A. Boel, R. Cartuyvels, Comparison of Vitek identification and antifungal susceptibility testing methods to DNA sequencing and Sensititre YeastOne antifungal testing, Med. Mycol. 49 (2011) 107–110.

[27] B. Posteraro, A. Ruggeri, E. De Carolis, R. Torelli, A. Vella, F. De Maio, et al., Comparative evaluation of BD Phoenix and vitek 2 systems for species identification of common and uncommon pathogenic yeasts, J. Clin. Microbiol. 51 (2013) 3841–3845.

[28] E.S. Spivak, K.E. Hanson, Candida auris: an emerging fungal pathogen, J. Clin. Microbiol. 56 (2018).

[29] N. Dhiman, L. Hall, S.L. Wohlfiel, S.P. Buckwalter, N.L. Wengenack, Performance and cost analysis of matrix-assisted laser desorption ionization-time of flight mass spectrometry for routine identification of yeast, J. Clin. Microbiol. 49 (2011) 1614–1616.

[30] D.A. Wilson, S. Young, K. Timm, S. Novak-Weekley, E.M. Marlowe, N. Madisen, et al., Multicenter evaluation of the Bruker MALDI biotyper CA system for the identification of clinically important bacteria and yeasts, Am. J. Clin. Pathol. 147 (2017) 623–631.

[31] J.T. Rudrik, M.K. Soehnlen, M.J. Perry, M.M. Sullivan, W. Reiter-Kintz, P.A. Lee, et al., Safety and accuracy of matrix-assisted laser desorption ionization-time of flight mass spectrometry for identification of highly pathogenic organisms, J. Clin. Microbiol. 55 (2017) 3513–3529.

[32] S. Rotcheewaphan, J.K. Lemon, U.U. Desai, C.M. Henderson, A.M. Zelazny, Rapid one-step protein extraction method for the identification of mycobacteria using MALDI-TOF MS, Diagn. Microbiol. Infect. Dis. 94 (2019) 355–360.

[33] C.B. Wilen, A.R. McMullen, C.A. Burnham, Comparison of sample preparation methods, instrumentation platforms, and contemporary commercial databases for identification of clinically relevant mycobacteria by matrix-assisted laser desorption ionization-time of flight mass spectrometry, J. Clin. Microbiol. 53 (2015) 2308–2315.

[34] L. Luo, W. Cao, W. Chen, R. Zhang, L. Jing, H. Chen, et al., Evaluation of the VITEK MS knowledge base version 3.0 for the identification of clinically relevant Mycobacterium species, Emerg. Microbes Infect. 7 (2018) 114.

[35] S.H. Kasperbauer, C.L. Daley, Mycobacterium chimaera infections related to the heater-cooler unit outbreak: a guide to diagnosis and management, Clin. Infect. Dis. 68 (2019) 1244–1250.

[36] R. Patel, A moldy application of MALDI: MALDI-ToF mass spectrometry for fungal identification, J. Fungi (Basel). 5 (2019).

[37] J. Rychert, E.S. Slechta, A.P. Barker, E. Miranda, N.E. Babady, Y.W. Tang, et al., Multicenter evaluation of the Vitek MS v3.0 system for the identification of filamentous fungi, J. Clin. Microbiol. 56 (2018).

[38] A.J. Loonen, A.R. Jansz, J. Stalpers, P.F. Wolffs, A.J. van den Brule, An evaluation of three processing methods and the effect of reduced culture times for faster direct identification of pathogens from BacT/ALERT blood cultures by MALDI-TOF MS, Eur. J. Clin. Microbiol. Infect. Dis. 31 (2012) 1575–1583.

[39] P.M. Juiz, M. Almela, C. Melcion, I. Campo, C. Esteban, C. Pitart, et al., A comparative study of two different methods of sample preparation for positive blood cultures for the rapid identification of bacteria using MALDI-TOF MS, Eur. J. Clin. Microbiol. Infect. Dis. 31 (2012) 1353–1358.

[40] C. Meex, F. Neuville, J. Descy, P. Huynen, M.P. Hayette, P. De Mol, et al., Direct identification of bacteria from BacT/ALERT anaerobic positive blood cultures by MALDI-TOF MS: MALDI Sepsityper kit versus an in-house saponin method for bacterial extraction, J. Med. Microbiol. 61 (2012) 1511–1516.

[41] S. Klein, S. Zimmermann, C. Kohler, A. Mischnik, W. Alle, K.A. Bode, Integration of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry in blood culture diagnostics: a fast and effective approach, J. Med. Microbiol. 61 (2012) 323–331.

[42] B.W. Buchan, K.M. Riebe, N.A. Ledeboer, Comparison of the MALDI Biotyper system using Sepsityper specimen processing to routine microbiological methods for identification of bacteria from positive blood culture bottles, J. Clin. Microbiol. 50 (2012) 346–352.

[43] L.G. Stevenson, S.K. Drake, P.R. Murray, Rapid identification of bacteria in positive blood culture broths by matrix-assisted laser desorption ionization-time of flight mass spectrometry, J. Clin. Microbiol. 48 (2010) 444–447.

[44] B. La Scola, D. Raoult, Direct identification of bacteria in positive blood culture bottles by matrix-assisted laser desorption ionisation time-of-flight mass spectrometry, PLoS One. 4 (2009) e8041.

[45] W. Moussaoui, B. Jaulhac, A.M. Hoffmann, B. Ludes, M. Kostrzewa, P. Riegel, et al., Matrix-assisted laser desorption ionization time-of-flight mass spectrometry identifies 90% of bacteria directly from blood culture vials, Clin. Microbiol. Infect. 16 (2010) 1631–1638.

[46] G. Prod'hom, A. Bizzini, C. Durussel, J. Bille, G. Greub, Matrix-assisted laser desorption ionization-time of flight mass spectrometry for direct bacterial identification from positive blood culture pellets, J. Clin. Microbiol. 48 (2010) 1481–1483.

[47] N.A. Ledeboer, B.K. Lopansri, N. Dhiman, R. Cavagnolo, K.C. Carroll, P. Granato, et al., Identification of gram-negative bacteria and genetic resistance determinants from positive blood culture Broths by use of the Verigene gram-negative blood culture multiplex microarray-based molecular assay, J. Clin. Microbiol. 53 (2015) 2460–2472.

[48] O. Altun, M. Almuhayawi, M. Ullberg, V. Ozenci, Clinical evaluation of the FilmArray blood culture identification panel in identification of bacteria and yeasts from positive blood culture bottles, J. Clin. Microbiol. 51 (2013) 4130–4136.

[49] M.L. Faron, B.W. Buchan, N.A. Ledeboer, Matrix-assisted laser desorption ionization-time of flight mass spectrometry for use with positive blood cultures: methodology, performance, and optimization, J. Clin. Microbiol. 55 (2017) 3328–3338.

[50] T. Spanu, B. Posteraro, B. Fiori, T. D'Inzeo, S. Campoli, A. Ruggeri, et al., Direct maldi-tof mass spectrometry assay of blood culture broths for rapid identification of *Candida* species causing bloodstream infections: an observational study in two large microbiology laboratories, J. Clin. Microbiol. 50 (2012) 176–179.

[51] M.L. Demarco, C.A. Burnham, Diafiltration MALDI-TOF mass spectrometry method for culture-independent detection and identification of pathogens directly from urine specimens, Am. J. Clin. Pathol. 141 (2014) 204–212.

[52] S. Montgomery, K. Roman, L. Ngyuen, A.M. Cardenas, J. Knox, A.P. Tomaras, et al., Prospective evaluation of light scatter technology paired with matrix-assisted laser desorption ionization-time of flight mass spectrometry for rapid diagnosis of urinary tract infections, J. Clin. Microbiol. 55 (2017) 1802–1811.

[53] J.F. Cohen, N. Bertille, R. Cohen, M. Chalumeau, Rapid antigen detection test for group A streptococcus in children with pharyngitis, Cochrane Database Syst. Rev. 7 (2016) CD010502.

[54] S.M. Babin, Y.H. Hsieh, R.E. Rothman, C.A. Gaydos, A meta-analysis of point-of-care laboratory tests in the diagnosis of novel 2009 swine-lineage pandemic influenza A (H1N1), Diagn. Microbiol. Infect. Dis. 69 (2011) 410–418.

[55] H. Chu, E.T. Lofgren, M.E. Halloran, P.F. Kuan, M. Hudgens, S.R. Cole, Performance of rapid influenza H1N1 diagnostic tests: a meta-analysis, Influenza Other Respir. Viruses. 6 (2012) 80–86.

[56] C. Chartrand, M.M. Leeflang, J. Minion, T. Brewer, M. Pai, Accuracy of rapid influenza diagnostic tests: a meta-analysis, Ann. Intern. Med. 156 (2012) 500–511.

[57] D.A. Green, K. StGeorge, Rapid antigen tests for influenza: rationale and significance of the FDA reclassification, J. Clin. Microbiol. 56 (2018).

[58] G.J. Berry, C.R. Miller, M.M. Prats, C. Marquez, O.O. Oladipo, M.J. Loeffelholz, et al., Comparison of the Alere i Strep A Test and the BD Veritor System in the detection of group A streptococcus and the hypothetical impact of results on antibiotic utilization, J. Clin. Microbiol. 56 (2018).

[59] K.G. Parker, S. Gandra, S. Matushek, K.G. Beavis, V. Tesic, A. Charnot-Katsikas, Comparison of 3 nucleic acid amplification tests and a rapid antigen test with culture for the detection of group A streptococci from throat swabs, J. Appl. Lab. Med. 4 (2019) 164–169.

[60] S. Young, P. Illescas, J. Nicasio, J.J. Sickler, Diagnostic accuracy of the real-time PCR cobas Liat Influenza A/B assay and the Alere i Influenza A&B NEAR isothermal nucleic acid amplification assay for the detection of influenza using adult nasopharyngeal specimens, J. Clin. Virol. 94 (2017) 86–90.

[61] M.J. Binnicker, M.J. Espy, C.L. Irish, E.A. Vetter, Direct detection of influenza A and B viruses in less than 20 minutes using a commercially available rapid PCR assay, J. Clin. Microbiol. 53 (2015) 2353–2354.

[62] J. Gibson, E.M. Schechter-Perkins, P. Mitchell, S. Mace, Y. Tian, K. Williams, et al., Multi-center evaluation of the cobas Liat Influenza A/B & RSV assay for rapid point of care diagnosis, J. Clin. Virol. 95 (2017) 5–9.

[63] T. Valentin, P. Kieslinger, E. Stelzl, B.I. Santner, A. Groselj-Strele, H.H. Kessler, et al., Prospective evaluation of three rapid molecular tests for seasonal influenza in patients presenting at an emergency unit, J. Clin. Virol. 111 (2019) 29–32.

[64] J.H. Chen, H.Y. Lam, C.C. Yip, V.C. Cheng, J.F. Chan, T.H. Leung, et al., Evaluation of the molecular Xpert Xpress Flu/RSV assay vs. Alere i Influenza A & B assay for rapid detection of influenza viruses, Diagn. Microbiol. Infect. Dis. 90 (2018) 177–180.

[65] J. Bell, A. Bonner, D.M. Cohen, R. Birkhahn, R. Yogev, W. Triner, et al., Multicenter clinical evaluation of the novel Alere i Influenza A&B isothermal nucleic acid amplification test, J. Clin. Virol. 61 (2014) 81–86.

[66] R.M. Peters, S.V. Schnee, J. Tabatabai, P. Schnitzler, J. Pfeil, Evaluation of Alere i RSV for rapid detection of respiratory syncytial virus in children hospitalized with acute respiratory tract infection, J. Clin. Microbiol. 55 (2017) 1032–1036.

[67] J. Merckx, R. Wali, I. Schiller, C. Caya, G.C. Gore, C. Chartrand, et al., Diagnostic accuracy of novel and traditional rapid tests for influenza infection compared with reverse transcriptase polymerase chain reaction: a systematic review and meta-analysis, Ann. Intern. Med. 167 (2017) 394–409.

[68] J.M. Martin, M. Green, K.A. Barbadora, E.R. Wald, Group A streptococci among school-aged children: clinical characteristics and the carrier state, Pediatrics. 114 (2004) 1212−1219.

[69] AAP, Red Book, Report of the Committee on Infectious Diseases, thirty-first ed., American Association of Pediatrics, 2018, pp 750, Itasca, IL, USA.

[70] E.B. Popowitch, S.S. O'Neill, M.B. Miller, Comparison of the Biofire FilmArray RP, Genmark eSensor RVP, Luminex xTAG RVPv1, and Luminex xTAG RVP fast multiplex assays for detection of respiratory viruses, J. Clin. Microbiol. 51 (2013) 1528−1533.

[71] N.E. Babady, M.R. England, K.L. Jurcic Smith, T. He, D.S. Wijetunge, Y.W. Tang, et al., Multicenter evaluation of the ePlex respiratory pathogen panel for the detection of viral and bacterial respiratory tract pathogens in nasopharyngeal swabs, J. Clin. Microbiol. 56 (2018).

[72] J.H.K. Chen, H.Y. Lam, C.C.Y. Yip, S.C.Y. Wong, J.F.W. Chan, E.S.K. Ma, et al., Clinical evaluation of the new high-throughput luminex NxTAG respiratory pathogen panel assay for multiplex respiratory pathogen detection, J. Clin. Microbiol. 54 (2016) 1820−1825.

[73] A.L. Leber, K. Everhart, J.A. Daly, A. Hopper, A. Harrington, P. Schreckenberger, et al., Multicenter evaluation of BioFire FilmArray respiratory panel 2 for detection of viruses and bacteria in nasopharyngeal swab samples, J. Clin. Microbiol. 56 (2018).

[74] K. Pabbaraju, S. Wong, K.L. Tokaryk, K. Fonseca, S.J. Drews, Comparison of the Luminex xTAG respiratory viral panel with xTAG respiratory viral panel fast for diagnosis of respiratory virus infections, J. Clin. Microbiol. 49 (2011) 1738−1744.

[75] U. Rappo, A.N. Schuetz, S.G. Jenkins, D.P. Calfee, T.J. Walsh, M.T. Wells, et al., Impact of early detection of respiratory viruses by multiplex PCR assay on clinical outcomes in adult patients, J. Clin. Microbiol. 54 (2016) 2096−2103.

[76] A. Subramony, P. Zachariah, A. Krones, S. Whittier, L. Saiman, Impact of multiplex polymerase chain reaction testing for respiratory pathogens on healthcare resource utilization for pediatric inpatients, J. Pediatr. 173 (2016) 196−201. e2.

[77] B.B. Rogers, P. Shankar, R.C. Jerris, D. Kotzbauer, E.J. Anderson, J.R. Watson, et al., Impact of a rapid respiratory panel test on patient outcomes, Arch. Pathol. Lab. Med. 139 (2015) 636−641.

[78] J.B. Mahony, G. Blackhouse, J. Babwah, M. Smieja, S. Buracond, S. Chong, et al., Cost analysis of multiplex PCR testing for diagnosing respiratory virus infections, J. Clin. Microbiol. 47 (2009) 2812−2817.

[79] A. Spina, K.G. Kerr, M. Cormican, F. Barbut, A. Eigentler, L. Zerva, et al., Spectrum of enteropathogens detected by the FilmArray GI Panel in a multicentre study of community-acquired gastroenteritis, Clin. Microbiol. Infect. 21 (2015) 719−728.

[80] S.N. Buss, A. Leber, K. Chapin, P.D. Fey, M.J. Bankowski, M.K. Jones, et al., Multicenter evaluation of the BioFire FilmArray gastrointestinal panel for etiologic diagnosis of infectious gastroenteritis, J. Clin. Microbiol. 53 (2015) 915−925.

[81] R. Khare, M.J. Espy, E. Cebelinski, D. Boxrud, L.M. Sloan, S.A. Cunningham, et al., Comparative evaluation of two commercial multiplex panels for detection of gastrointestinal pathogens by use of clinical stool specimens, J. Clin. Microbiol. 52 (2014) 3667−3673.

[82] E. Halligan, J. Edgeworth, K. Bisnauthsing, J. Bible, P. Cliff, E. Aarons, et al., Multiplex molecular testing for management of infectious gastroenteritis in a hospital setting: a comparative diagnostic and clinical utility study, Clin. Microbiol. Infect. 20 (2014) O460−O467.

[83] R.S. Huang, C.L. Johnson, L. Pritchard, R. Hepler, T.T. Ton, J.J. Dunn, Performance of the Verigene(R) enteric pathogens test, Biofire FilmArray gastrointestinal panel and Luminex xTAG(R) gastrointestinal pathogen panel for detection of common enteric pathogens, Diagn. Microbiol. Infect. Dis. 86 (2016) 336−339.

[84] E. Wessels, L.G. Rusman, M.J. van Bussel, E.C. Claas, Added value of multiplex Luminex Gastrointestinal Pathogen Panel (xTAG(R) GPP) testing in the diagnosis of infectious gastroenteritis, Clin. Microbiol. Infect. 20 (2014) O182−O187.

[85] S.D. Goldenberg, M. Bacelar, P. Brazier, K. Bisnauthsing, J.D. Edgeworth, A cost benefit analysis of the Luminex xTAG Gastrointestinal Pathogen Panel for detection of infectious gastroenteritis in hospitalised patients, J. Infect. 70 (2015) 504−511.

[86] K.H. Rand, E.E. Tremblay, M. Hoidal, L.B. Fisher, K.R. Grau, S. M. Karst, Multiplex gastrointestinal pathogen panels: implications for infection control, Diagn. Microbiol. Infect. Dis. 82 (2015) 154−157.

[87] C. Ward, K. Stocker, J. Begum, P. Wade, U. Ebrahimsa, S.D. Goldenberg, Performance evaluation of the Verigene(R) (Nanosphere) and FilmArray(R) (BioFire(R)) molecular assays for identification of causative organisms in bacterial bloodstream infections, Eur. J. Clin. Microbiol. Infect. Dis. 34 (2015) 487−496.

[88] M.M. Bhatti, S. Boonlayangoor, K.G. Beavis, V. Tesic, Evaluation of FilmArray and Verigene systems for rapid identification of positive blood cultures, J. Clin. Microbiol. 52 (2014) 3433−3436.

[89] T.D. Huang, E. Melnik, P. Bogaerts, S. Evrard, Y. Glupczynski, Evaluation of the ePlex blood culture identification panels for detection of pathogens in bloodstream infections, J. Clin. Microbiol. 57 (2019).

[90] M.J. Box, E.L. Sullivan, K.N. Ortwine, M.A. Parmenter, M.M. Quigley, L.M. Aguilar-Higgins, et al., Outcomes of rapid identification for gram-positive bacteremia in combination with antibiotic stewardship at a community-based hospital system, Pharmacotherapy. 35 (2015) 269−276.

[91] S.H. MacVane, F.S. Nolte, Benefits of adding a rapid PCR-based blood culture identification panel to an established antimicrobial stewardship program, J. Clin. Microbiol. 54 (2016) 2455−2463.

[92] K. Messacar, A.L. Hurst, J. Child, K. Campbell, C. Palmer, S. Hamilton, et al., Clinical impact and provider acceptability of real-time antimicrobial stewardship decision support for rapid diagnostics in children with positive blood culture results, J. Pediatric Infect. Dis. Soc. 6 (2017) 267−274.

[93] E.A. Neuner, A.M. Pallotta, S.W. Lam, D. Stowe, S.M. Gordon, G.W. Procop, et al., Experience with rapid microarray-based diagnostic technology and antimicrobial stewardship for patients with gram-positive bacteremia, Infect. Control. Hosp. Epidemiol. 37 (2016) 1361−1366.

[94] J. Pardo, K.P. Klinker, S.J. Borgert, B.M. Butler, P.G. Giglio, K. H. Rand, Clinical and economic impact of antimicrobial stewardship interventions with the FilmArray blood culture identification panel, Diagn. Microbiol. Infect. Dis. 84 (2016) 159−164.

[95] H. Suzuki, S. Hitomi, Y. Yaguchi, K. Tamai, A. Ueda, K. Kamata, et al., Prospective intervention study with a microarray-based, multiplexed, automated molecular diagnosis instrument (Verigene system) for the rapid diagnosis of bloodstream infections, and its impact on the clinical outcomes, J. Infect. Chemother. 21 (2015) 849–856.

[96] T. Walker, S. Dumadag, C.J. Lee, S.H. Lee, J.M. Bender, J. Cupo Abbott, et al., Clinical impact of laboratory implementation of verigene BC-GN microarray-based assay for detection of gram-negative bacteria in positive blood cultures, J. Clin. Microbiol. 54 (2016) 1789–1796.

[97] T.T. Timbrook, J.B. Morton, K.W. McConeghy, A.R. Caffrey, E. Mylonakis, K.L. LaPlante, The effect of molecular rapid diagnostic testing on clinical outcomes in bloodstream infections: a systematic review and meta-analysis, Clin. Infect. Dis. 64 (2017) 15–23.

[98] R. Banerjee, C.B. Teng, S.A. Cunningham, S.M. Ihde, J.M. Steckelberg, J.P. Moriarty, et al., Randomized trial of rapid multiplex polymerase chain reaction-based blood culture identification and susceptibility testing, Clin. Infect. Dis. 61 (2015) 1071–1080.

[99] A.L. Leber, K. Everhart, J.M. Balada-Llasat, J. Cullison, J. Daly, S. Holt, et al., Multicenter evaluation of BioFire FilmArray Meningitis/Encephalitis Panel for detection of bacteria, viruses, and yeast in cerebrospinal fluid specimens, J. Clin. Microbiol. 54 (2016) 2251–2261.

[100] E.H. Graf, M.V. Farquharson, A.M. Cardenas, Comparative evaluation of the FilmArray meningitis/encephalitis molecular panel in a pediatric population, Diagn. Microbiol. Infect. Dis. 87 (2017) 92–94.

[101] R.M. Liesman, A.P. Strasburg, A.K. Heitman, E.S. Theel, R. Patel, M.J. Binnicker, Evaluation of a commercial multiplex molecular panel for diagnosis of infectious meningitis and encephalitis, J. Clin. Microbiol. 56 (2018).

[102] P.O. Lewis, C.G. Lanier, P.D. Patel, W.D. Krolikowski, M.A. Krolikowski, False negative diagnostic errors with polymerase chain reaction for the detection of cryptococcal meningoencephalitis, Med. Mycol. (2019). Available from: https://doi.org/10.1093/mmy/myz064.

[103] G. DiDiodato, N. Bradbury, Cerebrospinal fluid analysis with the BioFire FilmArray Meningitis/Encephalitis Molecular Panel reduces length of hospital stay in patients with suspected central nervous system infections, Open. Forum Infect. Dis. 6 (2019) ofz119.

[104] K. Dack, S. Pankow, E. Ablah, R. Zackula, M. Assi, Contribution of the BioFire FilmArray Meningitis/Encephalitis Panel: assessing antimicrobial duration and length of stay, Kans. J. Med. 12 (2019) 1–3.

Self-assessment questions

1. What is the primary benefit of flocked vs. traditional swabs?
   - **a.** Flocked swabs are larger and allow for collection of a greater number of organisms
   - **b.** Flocked swabs allow a greater number of organisms to be released from the swab
   - **c.** Specimens collected using flocked swabs can be transported at any temperature
   - **d.** Flocked swabs are less expensive
2. Which of the following is not a component of current microbiology laboratory automation systems?
   - **a.** Specimen inoculation
   - **b.** Automated track
   - **c.** Anaerobic incubation
   - **d.** Digital imaging
3. Which of the following is the gold standard method for identification of microorganisms?
   - **a.** PCR
   - **b.** Biochemical methods
   - **c.** MALDI
   - **d.** DNA sequencing
4. Clinics in your hospital system perform rapid influenza diagnostic tests for the detection of influenza A and B viral antigens, which have a low negative predictive value (NPV) during influenza virus season. Which test performance characteristic is associated with the low NPV of influenza antigen testing?
   - **a.** High cross-reactivity with other viruses
   - **b.** Poor reproducibility
   - **c.** Poor sensitivity
   - **d.** Poor specificity
5. Several multiplex molecular panels are commercially available for syndromic-based microbiology diagnosis. Which of the following infections *cannot* be diagnosed by a syndromic-based test?
   - **a.** Central nervous system infection
   - **b.** Gastrointestinal infection
   - **c.** Prosthetic joint infection
   - **d.** Upper respiratory infection

Answers

1. b
2. c, anaerobic incubation
3. d, DNA sequencing
4. c, Poor sensitivity
5. c, Prosthetic joint infection

Index

Note: Page numbers followed by "*f*" and "*t*" refer to figures and tables, respectively.

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z